The Missing Link: Dedicated End-to-End 10Gbps Optical Lightpaths for Clusters, Grids, and Clouds

Invited Keynote Presentation

11th IEEE/ACM International Symposium on Cluster, Cloud, and Grid Computing

Newport Beach, CA

May 24, 2011

Dr. Larry Smarr

Director, California Institute for Telecommunications and Information Technology

Harry E. Gruber Professor,

Dept. of Computer Science and Engineering

Jacobs School of Engineering, UCSD

Follow me on Twitter: lsmarr

Abstract

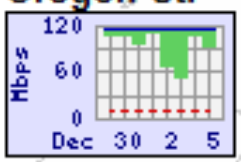
Today we are living in a data-dominated world where distributed scientific instruments, as well as clusters, generate terabytes to petabytes of data which are stored increasingly in specialized campus facilities or in the Cloud. It was in response to this challenge that the NSF funded the OptIPuter project to research how user-controlled 10Gbps dedicated lightpaths (or "lambdas") could transform the Grid into a LambdaGrid. This provides direct access to global data repositories, scientific instruments, and computational resources from "OptIPortals," PC clusters which provide scalable visualization, computing, and storage in the user's campus laboratory. The use of dedicated lightpaths over fiber optic cables enables individual researchers to experience "clear channel" 10,000 megabits/sec, 100-1000 times faster than over today's shared Internet-a critical capability for data-intensive science. The seven-year OptIPuter computer science research project is now over, but it stimulated a national and global build-out of dedicated fiber optic networks. U.S. universities now have access to high bandwidth lambdas through the National LambdaRail, Internet2's WaveCo, and the Global Lambda Integrated Facility. A few pioneering campuses are now building on-campus lightpaths to connect the data-intensive researchers, data generators, and vast storage systems to each other on campus, as well as to the national network campus gateways. I will give examples of the application use of this emerging high performance cyberinfrastructure in genomics, ocean observatories, radio astronomy, and cosmology.

Large Data Challenge: Average Throughput to End User on Shared Internet is 10-100 Mbps

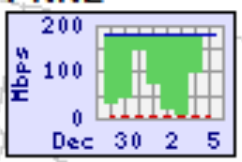


Tested
January 2011

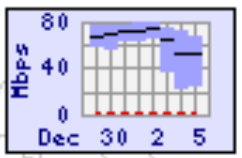
Oregon St.



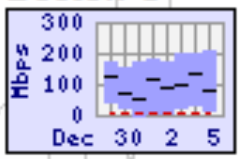
PNNL



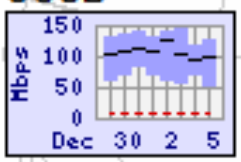
Toronto



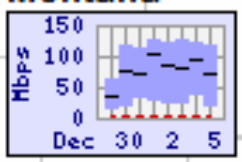
Boston U



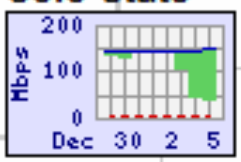
UCSB



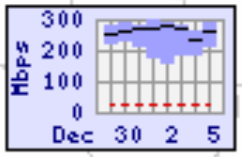
Montana



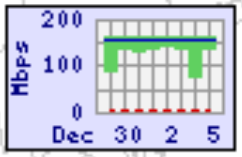
Colo State



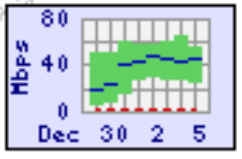
U. Wisconsin



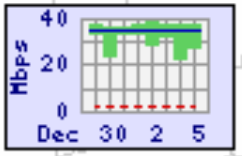
Penn State



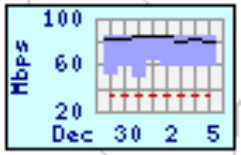
SUNY-SB



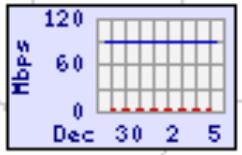
UIUC



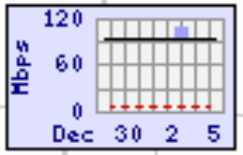
JPL-MISR



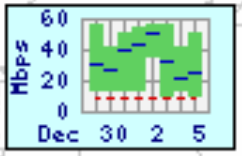
JPL-ASTER-IST



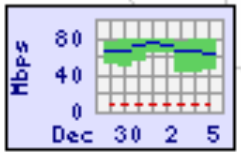
NCAR



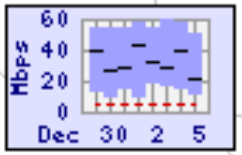
UAH-GHRC



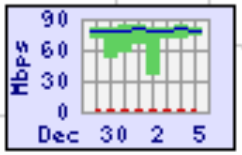
UCSD



Arizona



LANL



Transferring 1 TB:
--50 Mbps = 2 Days
--10 Gbps = 15 Minutes

http://ensight.eos.nasa.gov/Missions/terra/index.shtml

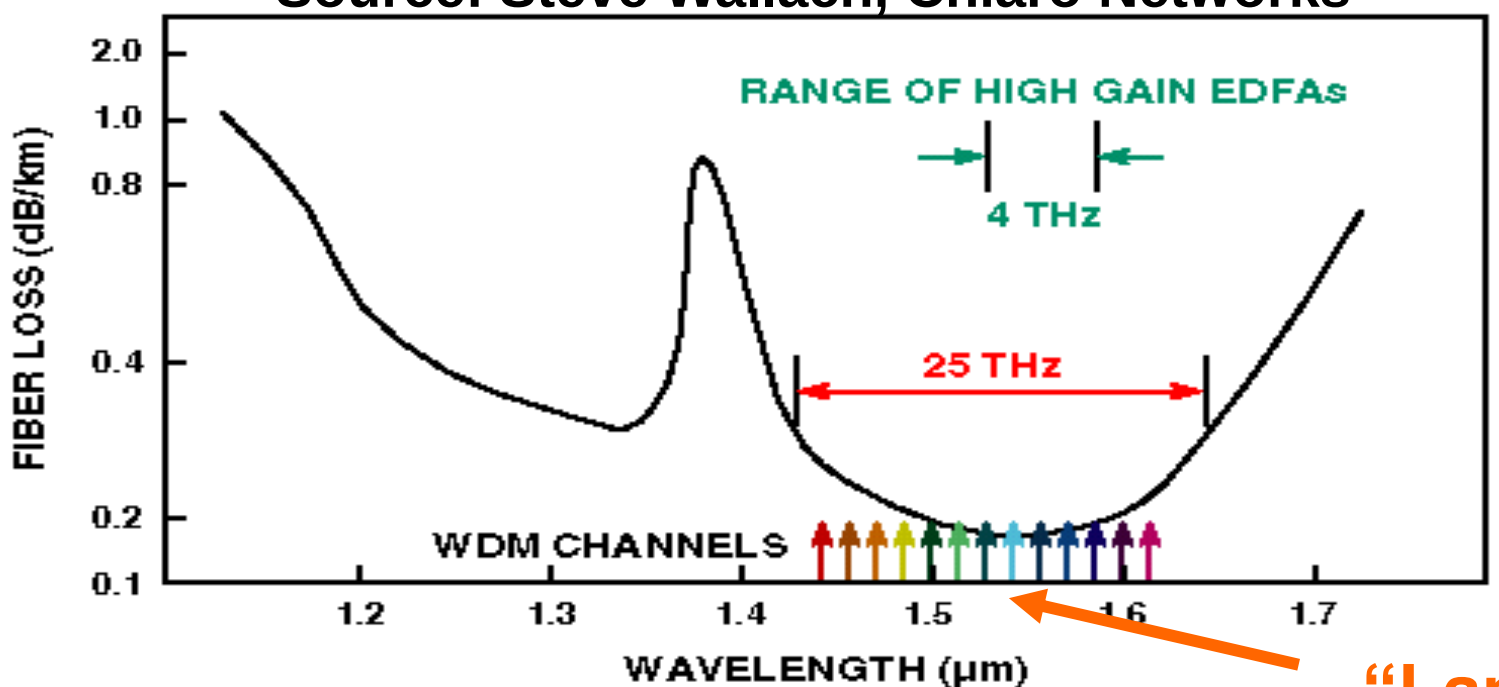
OptIPuter Solution:
Give Dedicated Optical Channels to Data-Intensive Users

WAVELENGTH DIVISION MULTIPLEXING (WDM)

10 Gbps per User >100x Shared Internet Throughput

- EXPLOITS
  - ENORMOUS BANDWIDTH OF SILICA FIBER
  - HIGH-GAIN WIDEBAND OPTICAL AMPLIFIERS

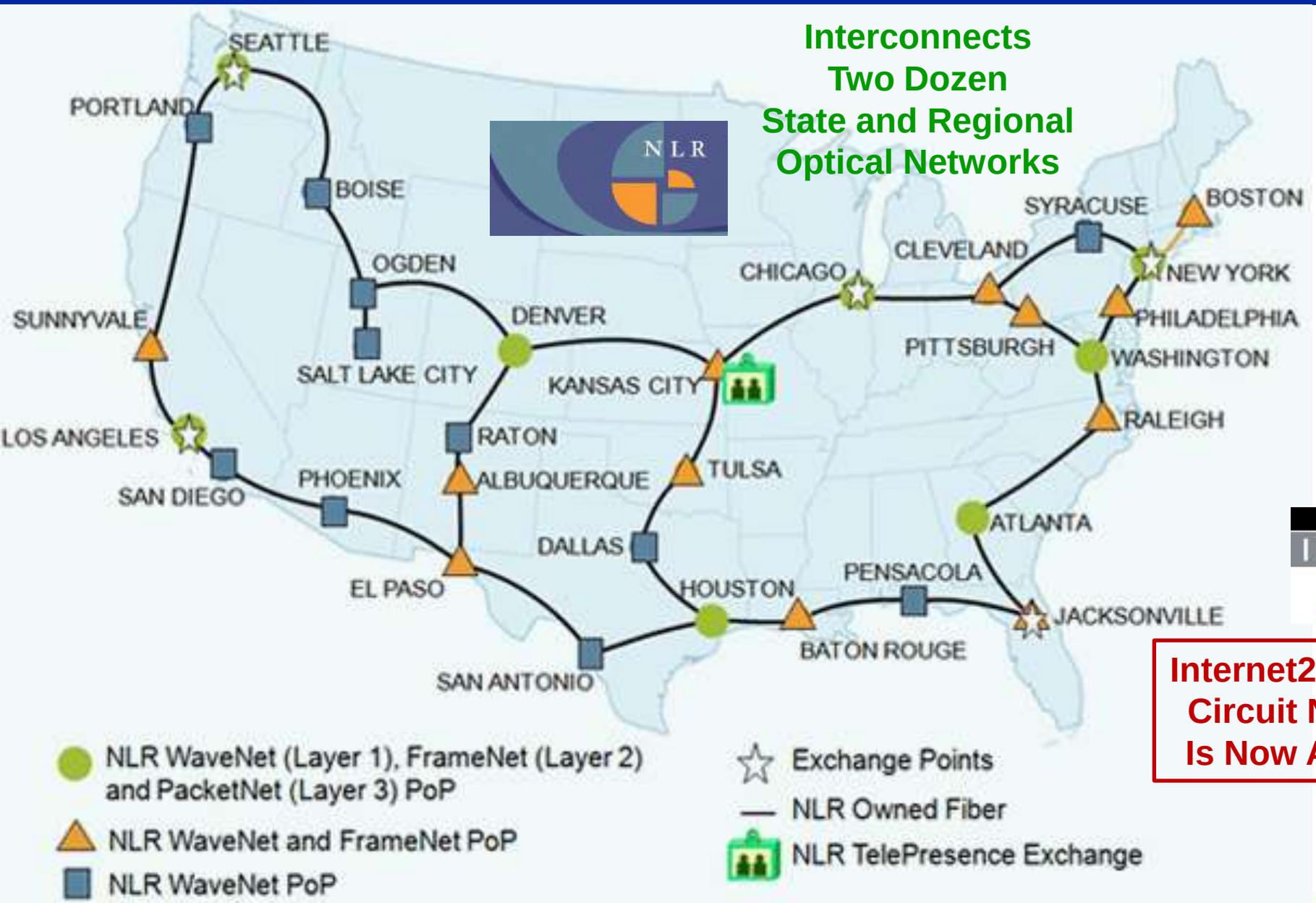
Source: Steve Wallach, Chiaro Networks



"Lambdas"

Parallel Lambdas are Driving Optical Networking The Way Parallel Processors Drove 1990s Computing

Dedicated 10Gbps Lightpaths Tie Together State and Regional Fiber Infrastructure

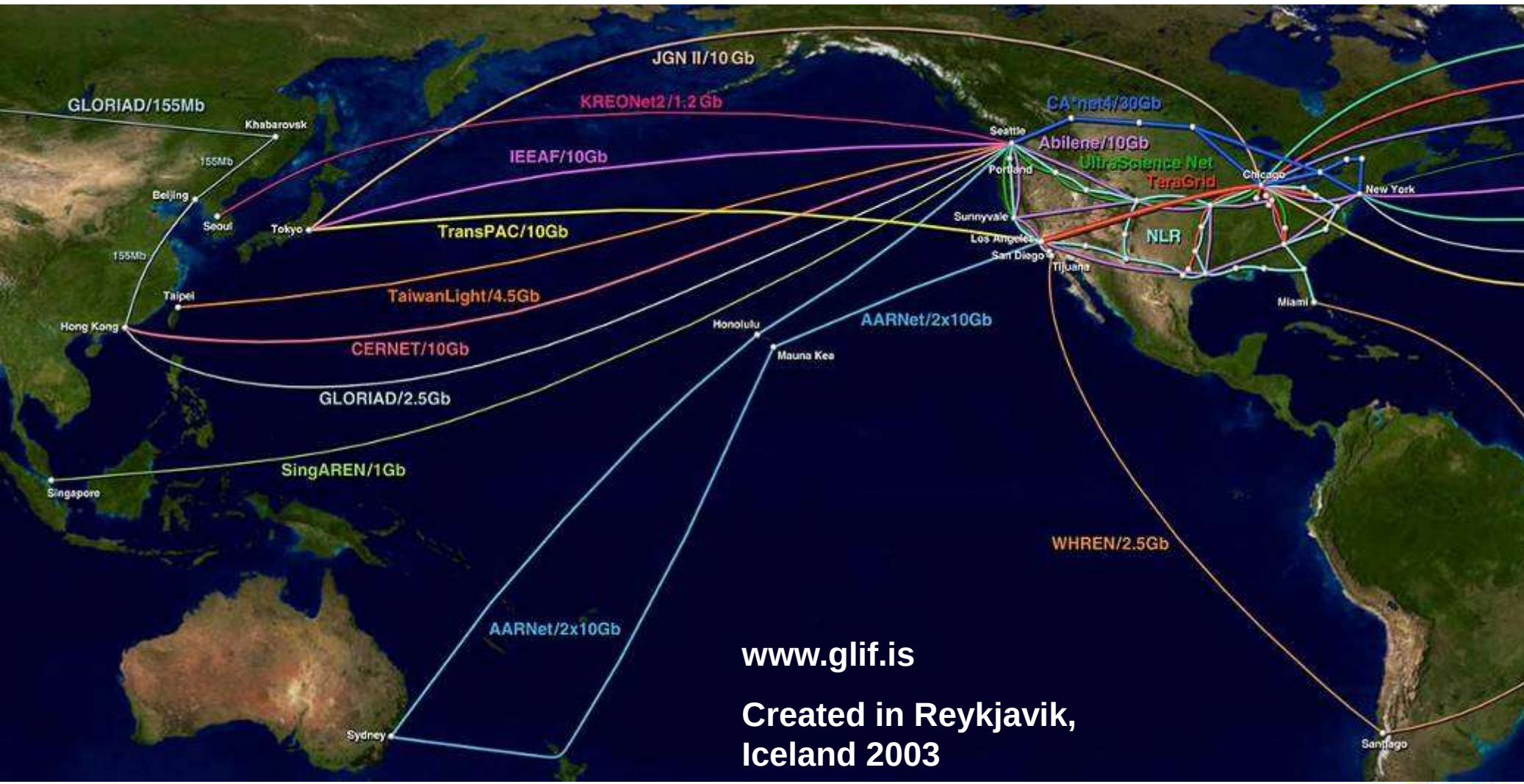


INTERNET 2

Internet2 WaveCo Circuit Network Is Now Available

The Global Lambda Integrated Facility--
Creating a Planetary-Scale High Bandwidth Collaboratory

Research Innovation Labs Linked by 10G Dedicated Lambdas



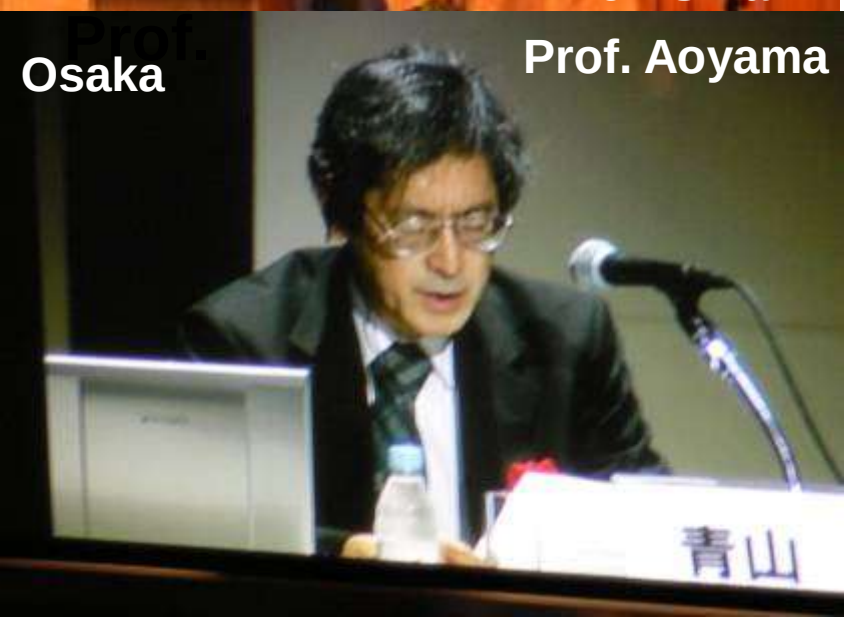
www.glif.is

Created in Reykjavik, Iceland 2003

Visualization courtesy of
Bob Patterson, NCSA.

High Resolution Uncompressed HD Streams Require Multi-Gigabit/s Lambdas





Telepresence Using Uncompressed 1.5 Gbps HDTV Streaming Over IP on Fiber Optics-- 75x Home Cable "HDTV" Bandwidth!



"I can see every hair on your head!"—Prof. Aoyama

Source: U Washington Research Channel

Borderless Collaboration
Between Global University Research Centers at 10Gbps

Maxine Brown, Tom DeFanti, Co-Chairs

iGrid 2005

THE GLOBAL LAMBDA INTEGRATED FACILITY

www.igrid2005.org

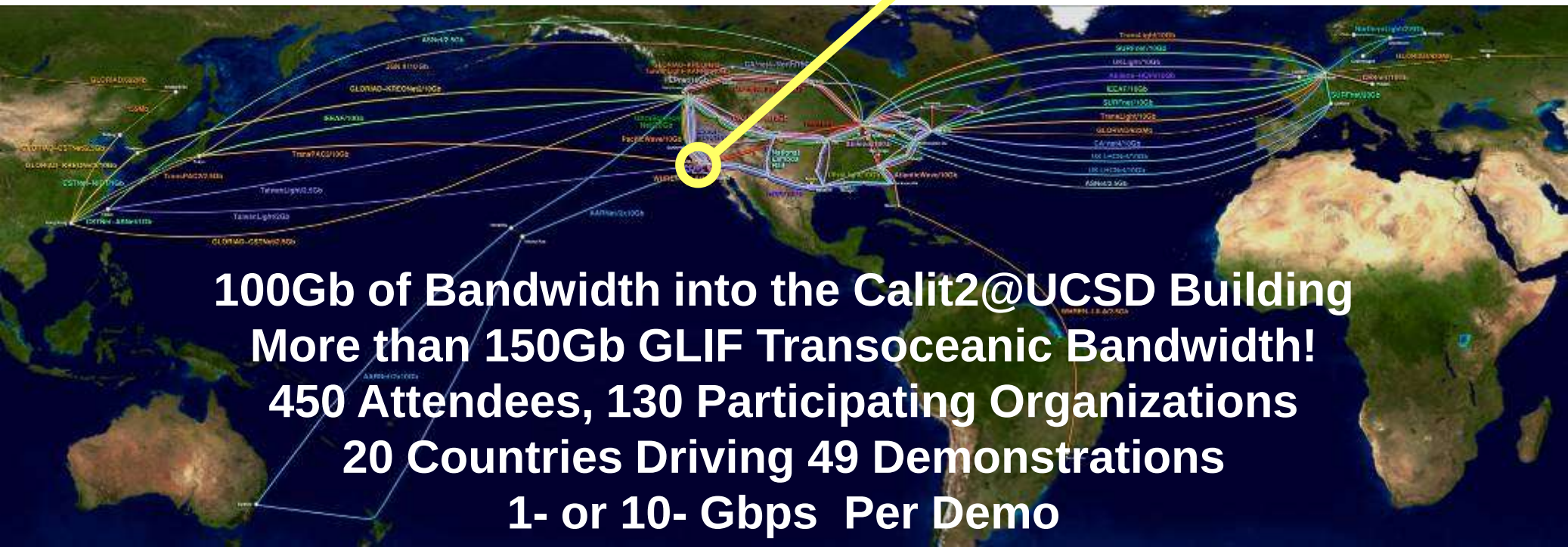


the UC Discovery Grant

September 26-30, 2005

Calit2 @ University of California, San Diego

California Institute for Telecommunications and Information Technology



100Gb of Bandwidth into the Calit2@UCSD Building

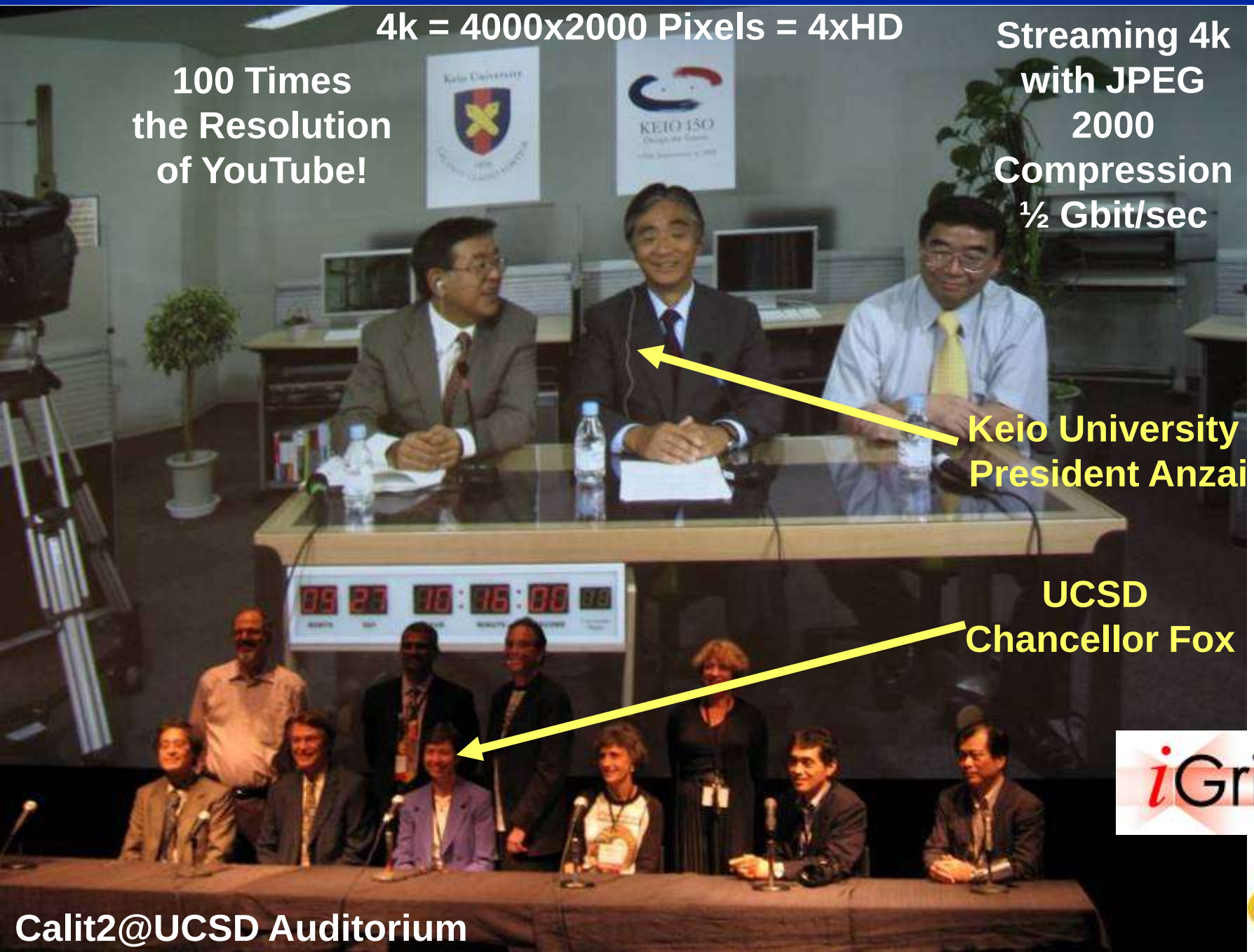
More than 150Gb GLIF Transoceanic Bandwidth!

450 Attendees, 130 Participating Organizations

20 Countries Driving 49 Demonstrations

1- or 10- Gbps Per Demo

Telepresence Meeting
Using Digital Cinema 4k Streams



CineGrid

Lays Technical Basis for Global Digital Cinema

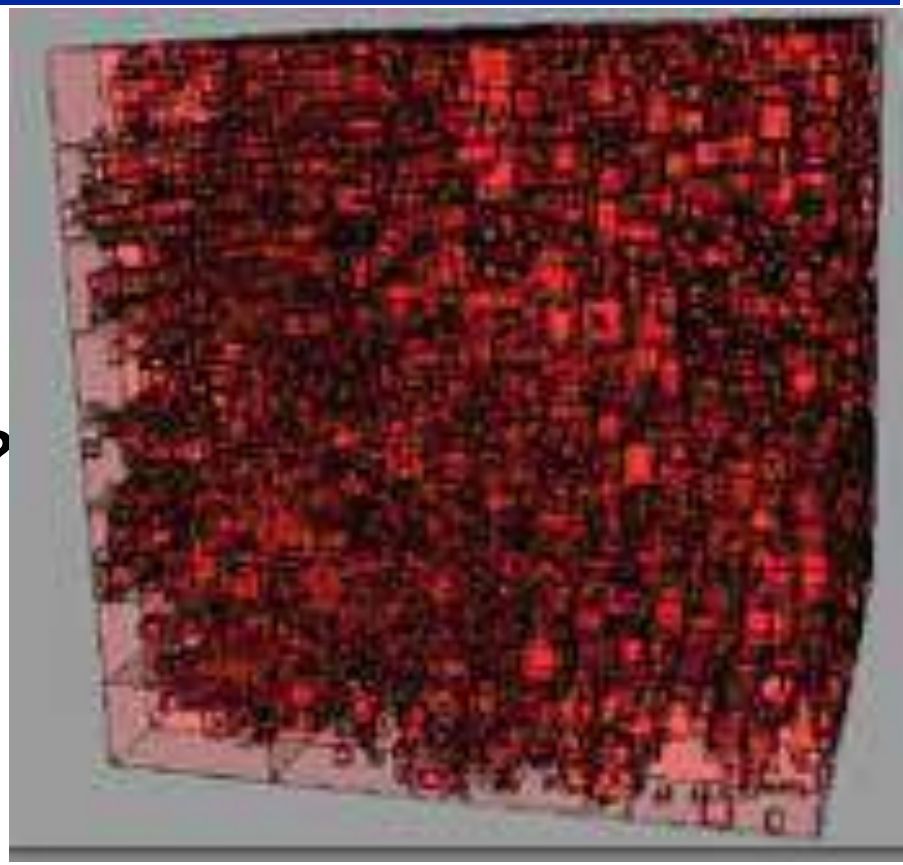
Sony
NTT
SGI

iGrid 2005

it²

iGrid Lambda High Performance Computing Services: Distributing AMR Cosmology Simulations

- **Uses ENZO Computational Cosmology Code**
  - **Grid-Based Adaptive Mesh Refinement Simulation Code**
  - **Developed by Mike Norman, UCSD**
- **Can One Distribute the Computing?**
  - **iGrid2005 to Chicago to Amsterdam**
- **Distributing Code Using Layer 3 Routers Fails**
- **Instead Using Layer 2, Essentially Same Performance as Running on Single Supercomputer**
  - **Using Dynamic Lightpath Provisioning**



Source: Joe Mambretti, Northwestern U

iGrid Lambda Control Services: Transform Batch to Real-Time Global e-Very Long Baseline Interferometry

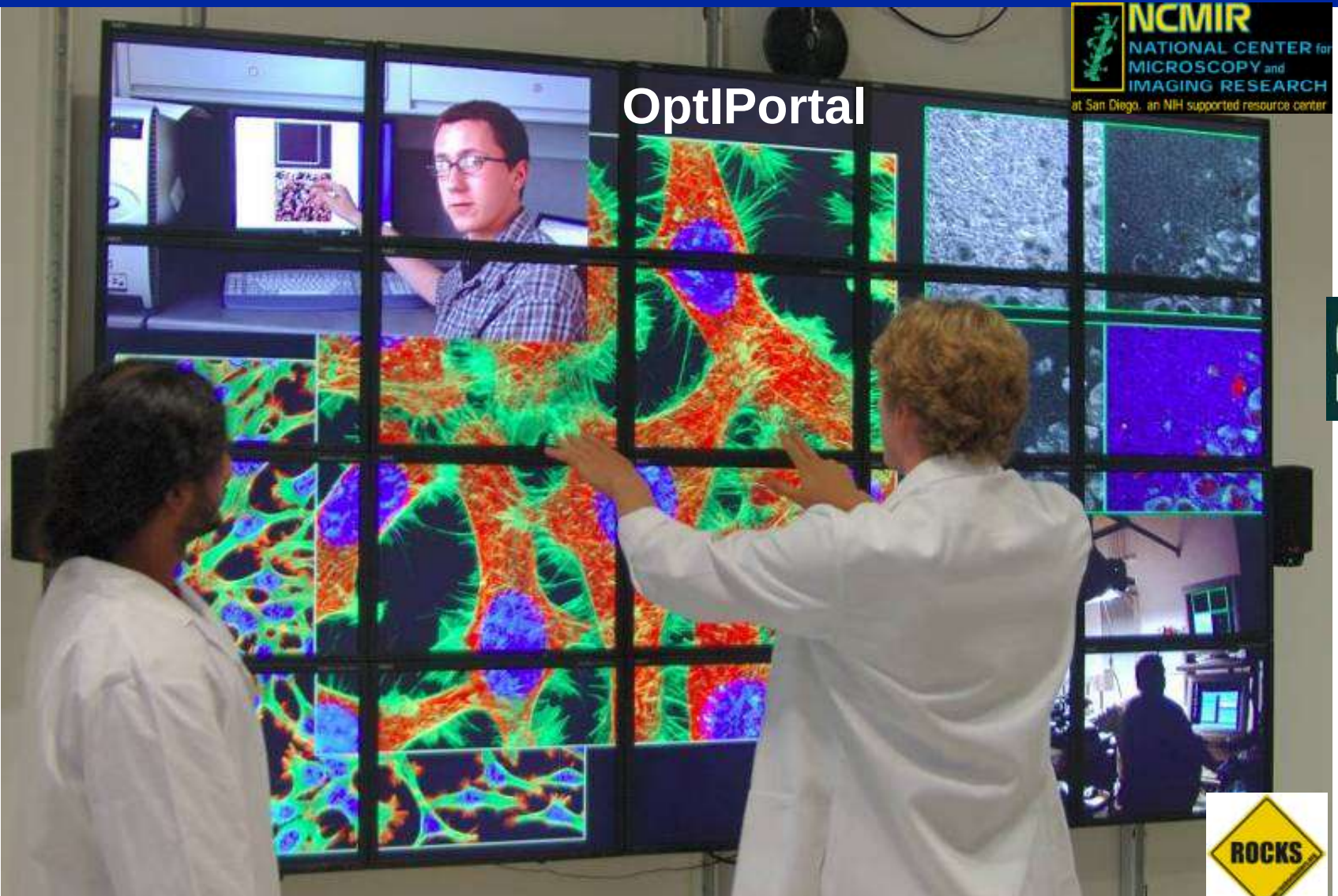


- **Goal: Real-Time VLBI Radio Telescope Data Correlation**
- **Achieved 512Mb Transfers from USA and Sweden to MIT**
- **Results Streamed to iGrid2005 in San Diego**

Optical Connections Dynamically Managed Using the DRAGON Control Plane and Internet2 HOPI Network

Source: Jerry Sobieski, DRAGON

The OptIPuter Project: Creating High Resolution Portals Over Dedicated Optical Channels to Global Science Data



NCMIR NATIONAL CENTER for MICROSCOPY and IMAGING RESEARCH at San Diego, an NIH supported resource center

Scalable Adaptive Graphics Environment (SAGE)

SAGE

evl electronic visualization laboratory

NSF

ROCKS

Picture Source: Mark Ellisman, David Lee, Jason Leigh

OptIPuter

Calit2 (UCSD, UCI), SDSC, and UIC Leads—Larry Smarr PI

Univ. Partners: NCSA, USC, SDSU, NW, TA&M, UvA, SARA, KISTI, AIST

Industry: IBM, Sun, Telcordia, Chiaro, Calient, Glimmerglass, Lucent

it²

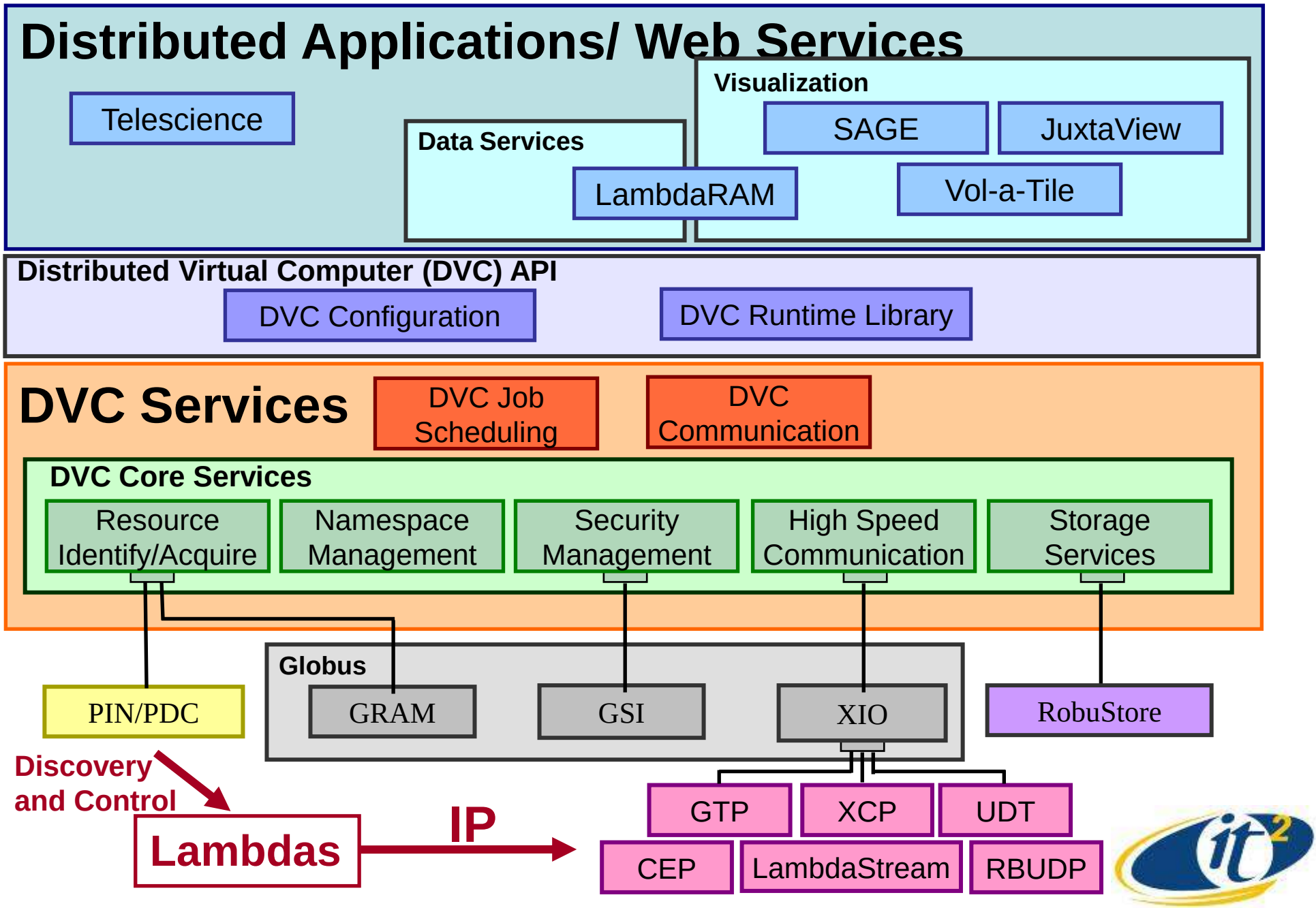
What is the OptIPuter?

- Applications Drivers → Interactive Analysis of Large Data Sets
- OptIPuter Nodes → Scalable PC Clusters with Graphics Cards
- IP over Lambda Connectivity →Predictable Backplane
- Open Source LambdaGrid Middleware→ Network is Reservable
- Data Retrieval and Mining → Lambda Attached Data Servers
- High Defn. Vis., Collab. SW → High Performance Collaboratory

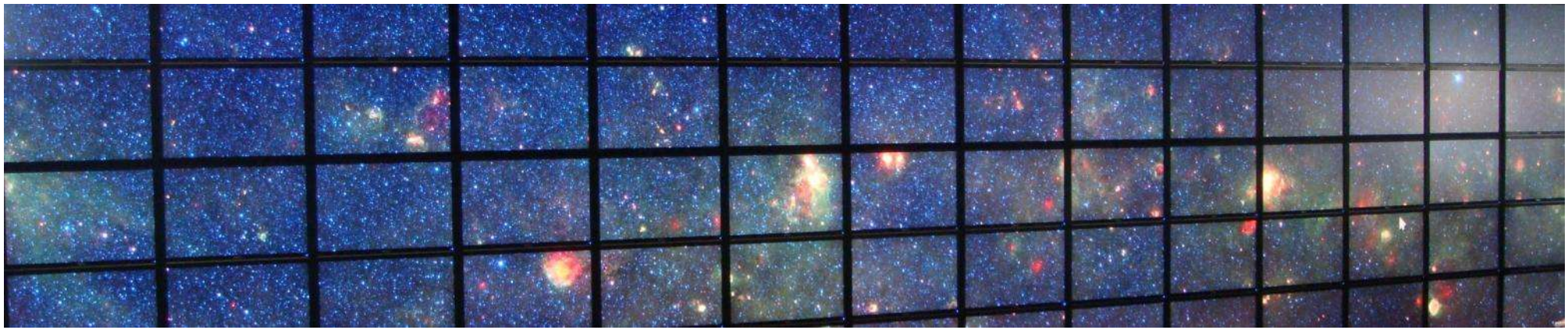
www.optiputer.net

See Nov 2003 Communications of the ACM for Articles on OptIPuter Technologies

OptIPuter Software Architecture--a Service-Oriented Architecture Integrating Lambdas Into the Grid



OptIPortals Scale to 1/3 Billion Pixels Enabling Viewing of Very Large Images or Many Simultaneous Images



Spitzer Space Telescope (Infrared)



NASA Earth Satellite Images Bushfires October 2007 San Diego

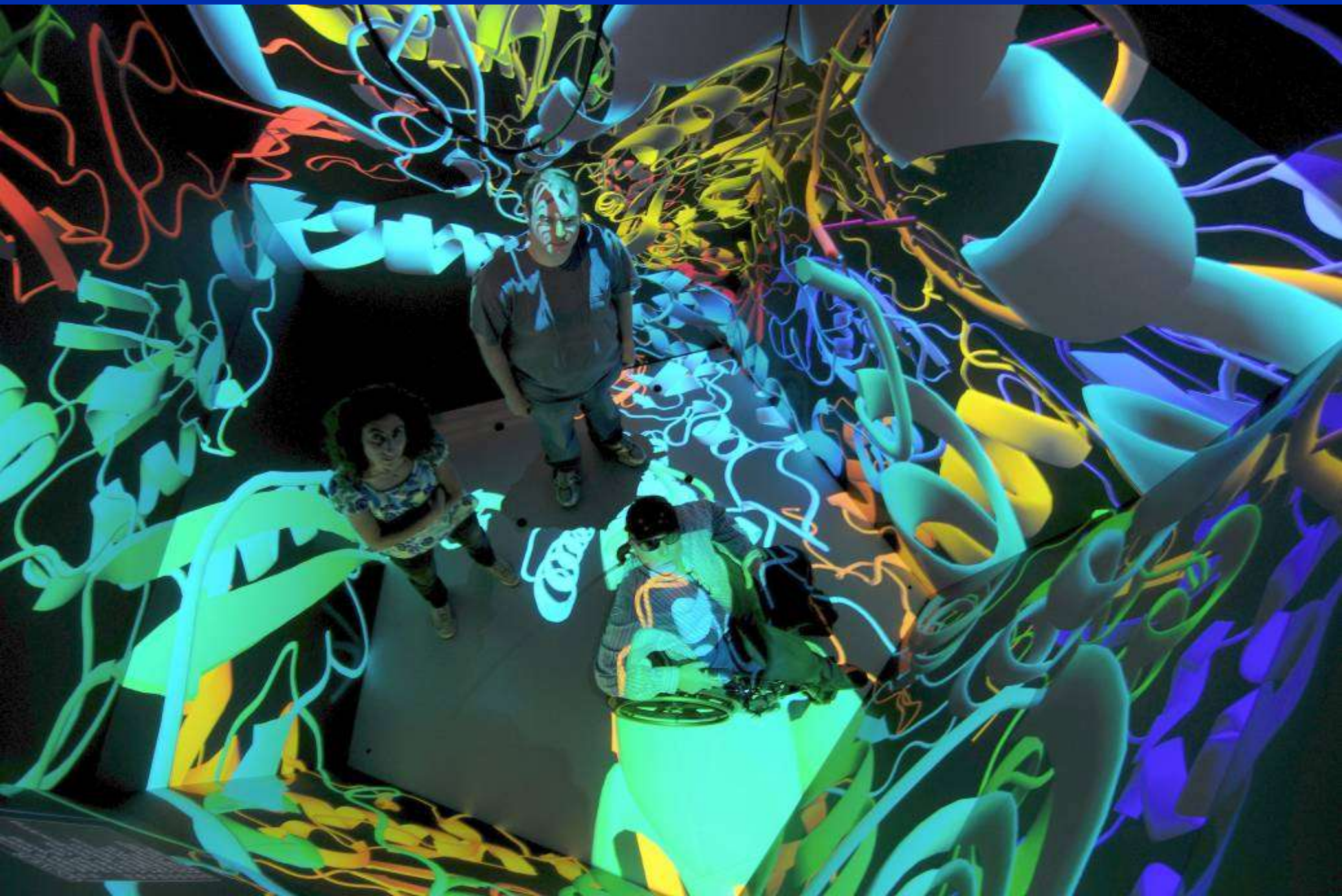
Source: Falko Kuester, Calit2@UCSD

The Latest OptIPuter Innovation: Quickly Deployable Nearly Seamless OptIPortables



45 minute setup, 15 minute tear-down with two people (possible with one)

Calit2 3D Immersive StarCAVE OptIPortal



3D Stereo Head Tracked OptIPortal: NexCAVE

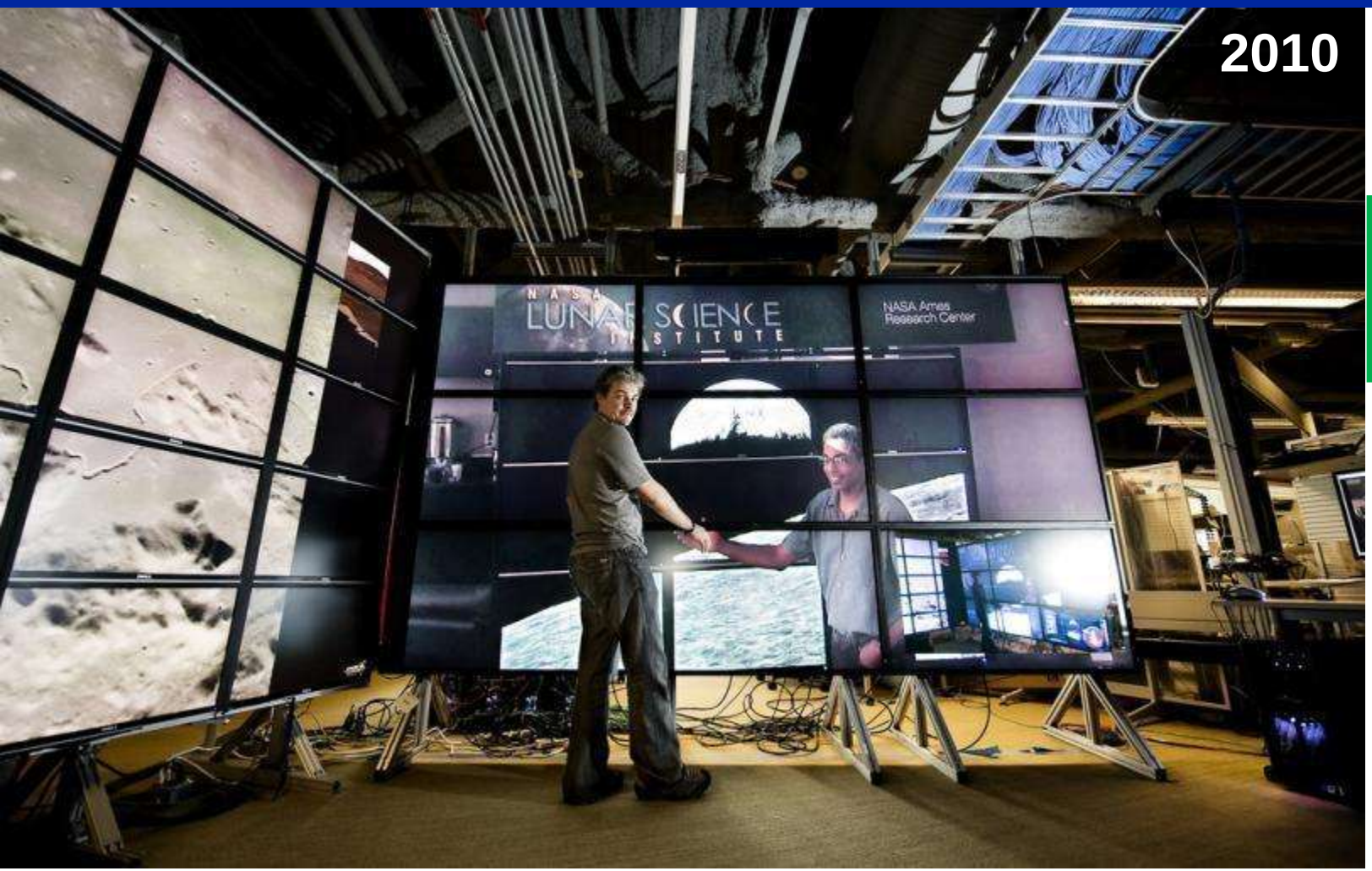


Array of JVC HDTV 3D LCD Screens
KAUST NexCAVE = 22.5MPixels

www.calit2.net/newsroom/article.php?id=1584

Source: Tom DeFanti, Calit2@UCSD

High Definition Video Connected OptIPortals: Virtual Working Spaces for Data Intensive Research



NASA Supports Two Virtual Institutes

CGLX

LifeSize HD

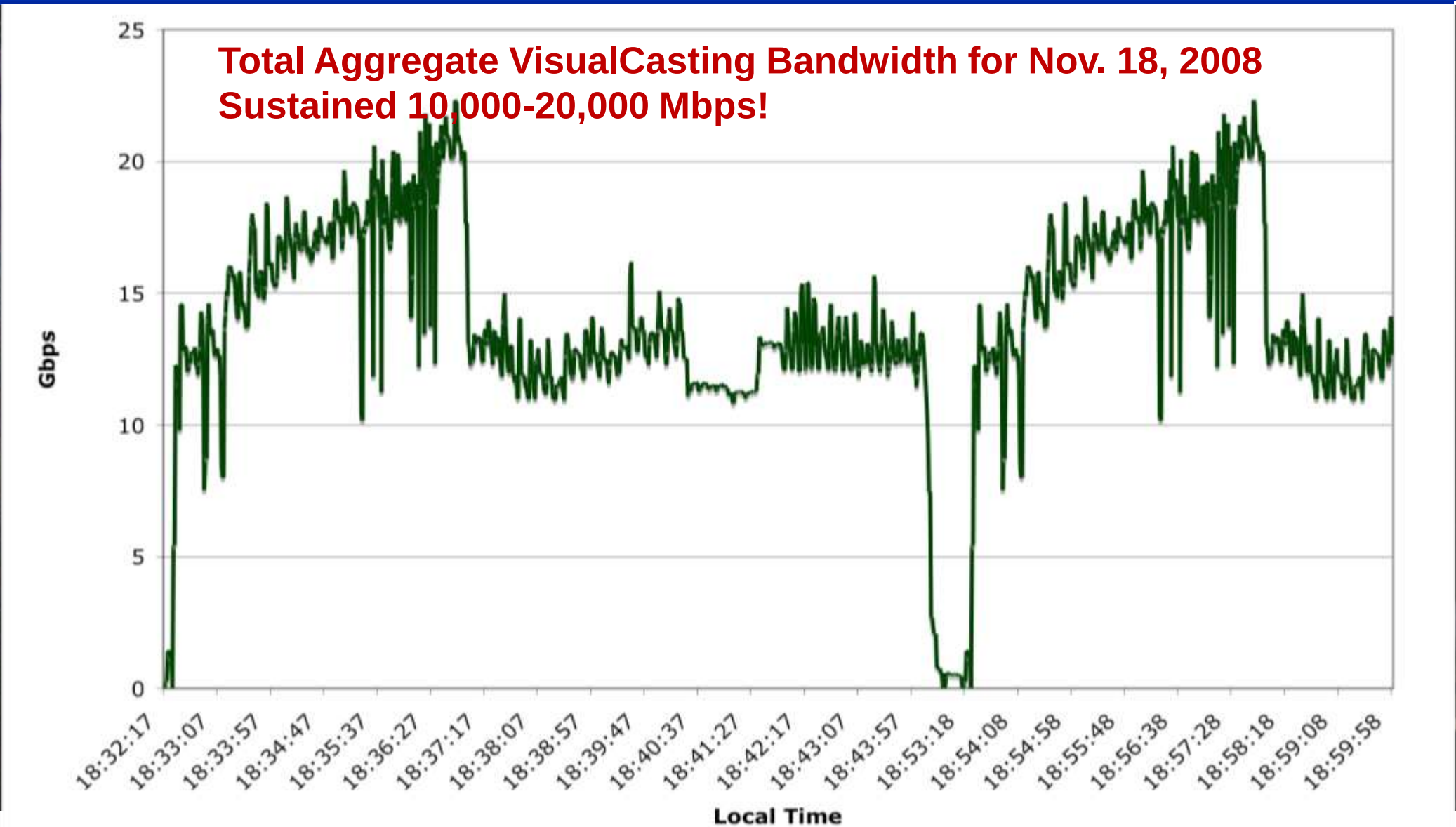
Calit2@UCSD 10Gbps Link to NASA Ames Lunar Science Institute, Mountain View, CA

OptIPuter

Source: Falko Kuester, Kai Doerr Calit2; Michael Sims, Larry Edwards, Estelle Dodson NASA

it²

EVL's SAGE OptIPortal VisualCasting Multi-Site OptIPuter Collaboratory



Source: Jason Leigh, Luc Renambot, EVL, UI Chicago

Using Supernetworks to Couple End User's OptIPortal to Remote Supercomputers and Visualization Servers

Source: Mike Norman, Rick Wagner, SDSC

Argonne NL

DOE Eureka
100 Dual Quad Core Xeon Servers
200 NVIDIA Quadro FX GPUs in 50 Quadro Plex S4 1U enclosures
3.2 TB RAM



SDSC

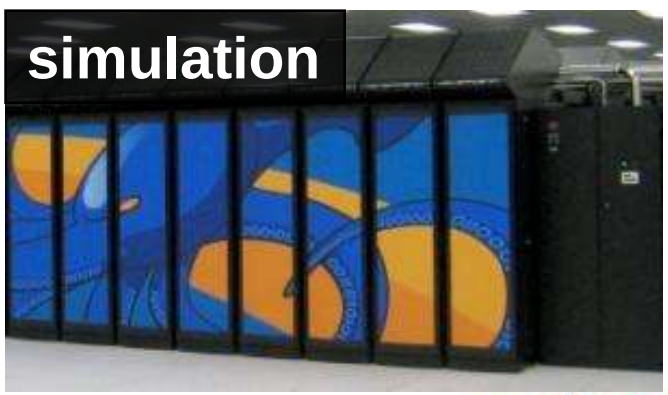


ESnet
10 Gb/s fiber optic network

NICS ORNL

Calit2/SDSC OptIPortal1
20 30" (2560 x 1600 pixel) LCD panels
10 NVIDIA Quadro FX 4600 graphics cards > 80 megapixels
10 Gb/s network throughout

NSF TeraGrid Kraken
Cray XT5
8,256 Compute Nodes
99,072 Compute Cores
129 TB RAM



SDSC

***ANL * Calit2 * LBNL * NICS * ORNL * SDSC**

it²

National-Scale Interactive Remote Rendering of Large Datasets

SDSC



OptIPortal (40M pixels LCDs)
10 NVIDIA FX 4600 Cards
10 Gb/s Network Throughout

ESnet

Science Data Network (SDN)
> 10 Gb/s Fiber Optic Network
Dynamic VLANs Configured
Using OSCARS

ALCF



Eureka
100 Dual Quad Core Xeon Servers
200 NVIDIA FX GPUs
3.2 TB RAM

Interactive Remote Rendering

Real-Time Volume Rendering Streamed from ANL to SDSC

Last Year

High-Resolution (4K+, 15+ FPS)—But:
- Command-Line Driven
- Fixed Color Maps, Transfer Functions
- Slow Exploration of Data

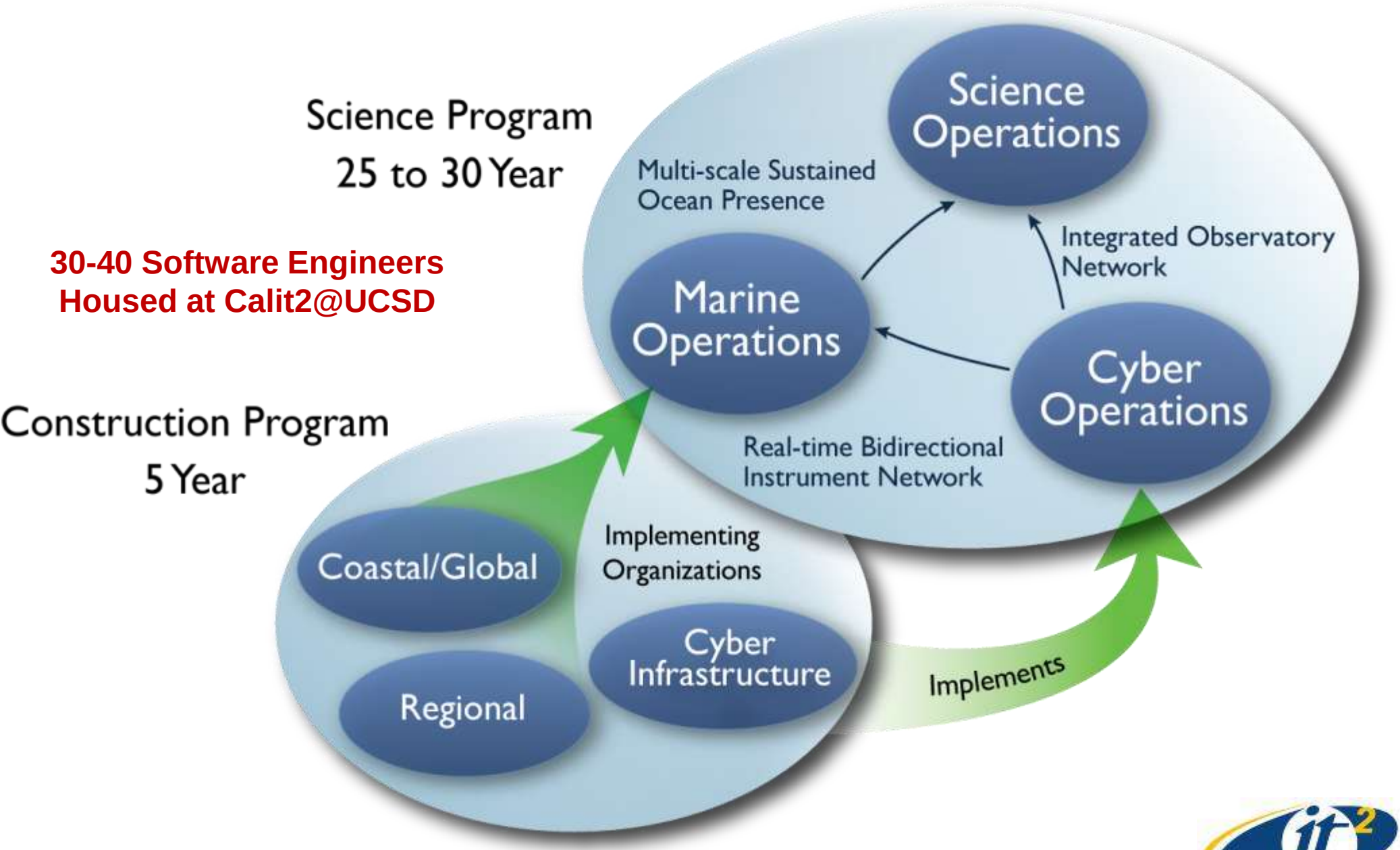
Now

Driven by a Simple Web GUI:
- Rotate, Pan, Zoom
- GUI Works from Most Browsers
- Manipulate Colors and Opacity
- Fast Renderer Response Time

Source: Rick Wagner, SDSC

NSF OOI is a $400M Program -OOI CI is $34M Part of This

OCEAN OBSERVATORIES INITIATIVE

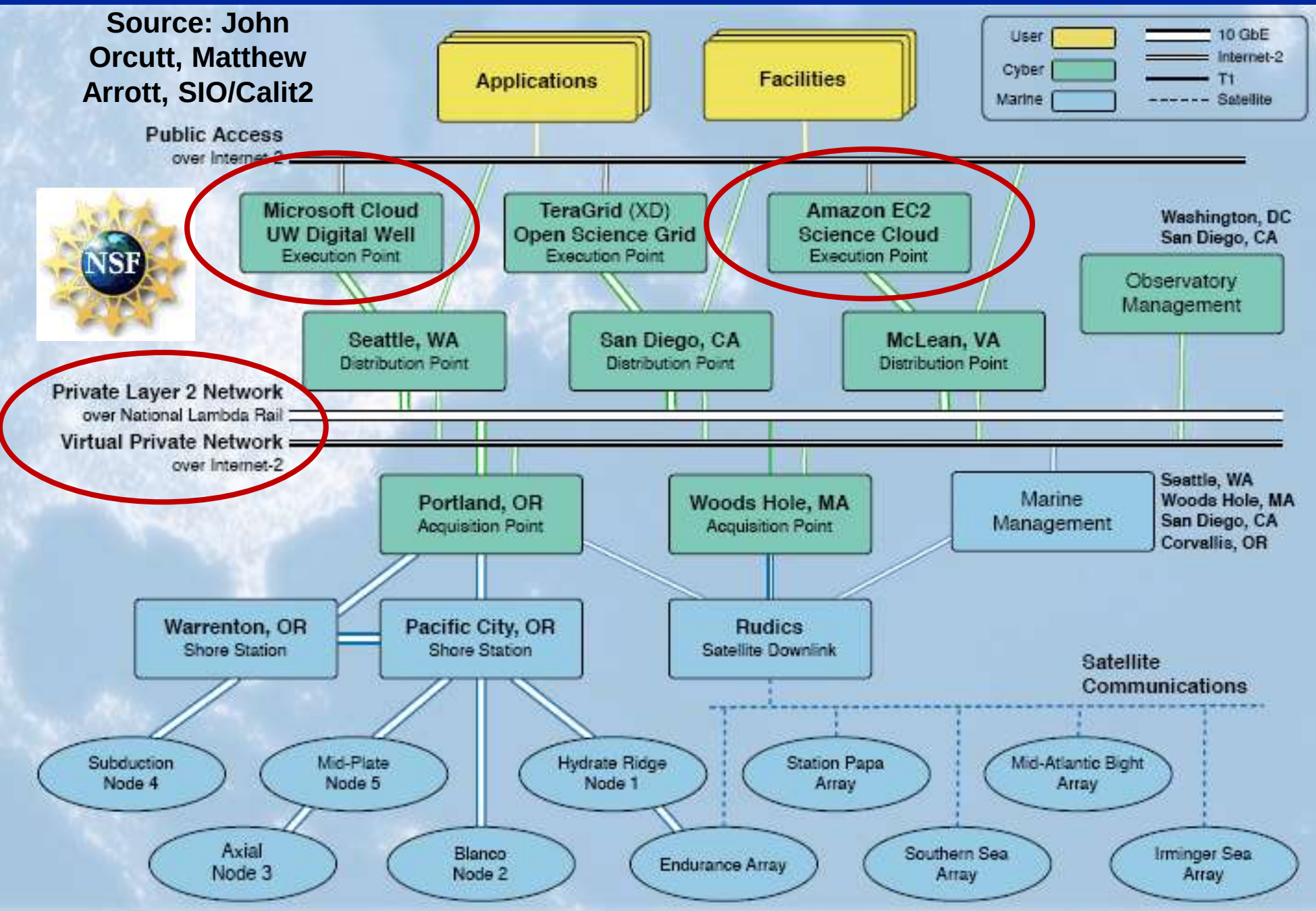


30-40 Software Engineers Housed at Calit2@UCSD

Source: Matthew Arrott, Calit2 Program Manager for OOI CI

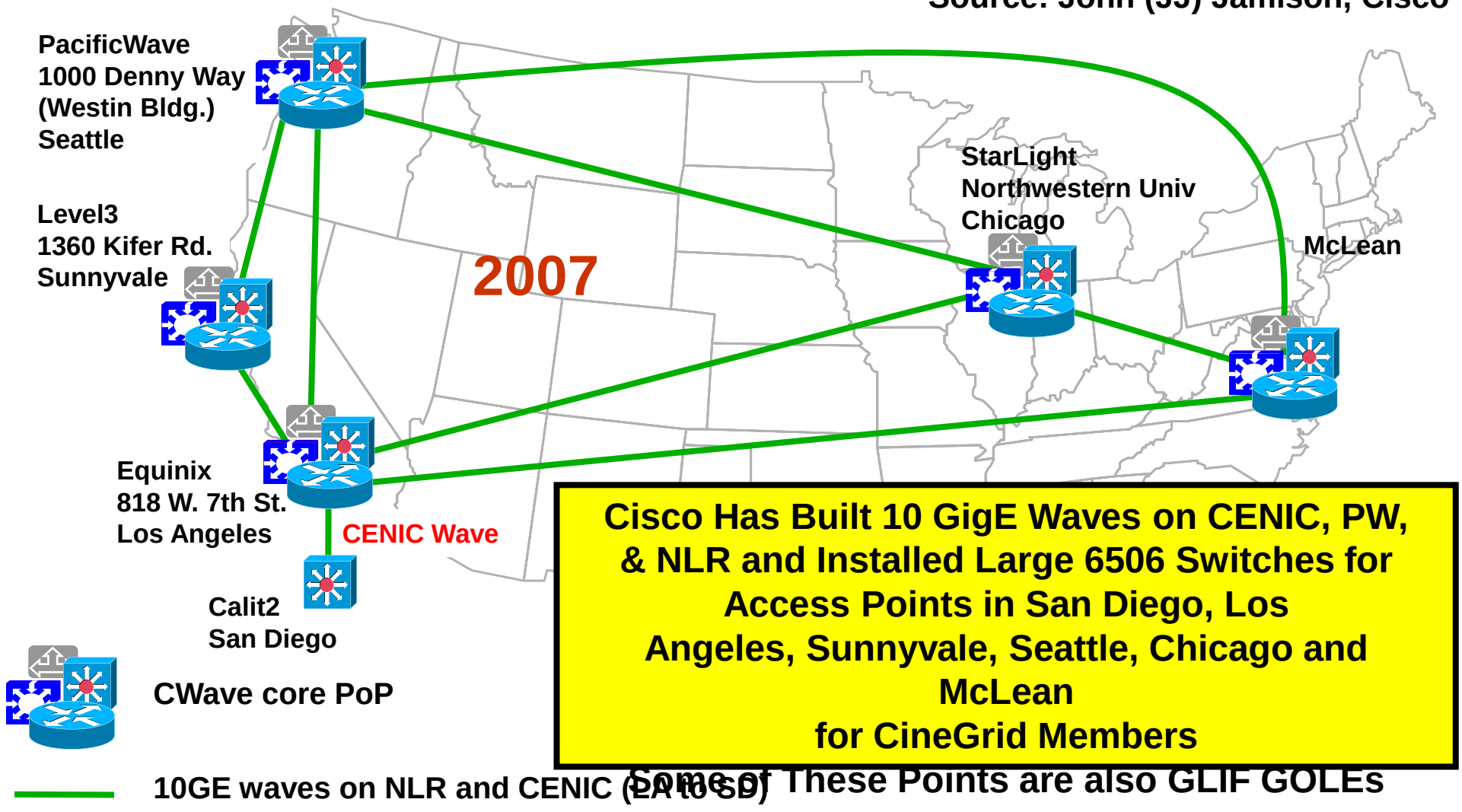
OOI CI
is Built on NLR/I2 Optical Infrastructure

OCEAN OBSERVATORIES INITIATIVE



Cisco CWave for CineGrid: A New Cyberinfrastructure for High Resolution Media Streaming*

Source: John (JJ) Jamison, Cisco



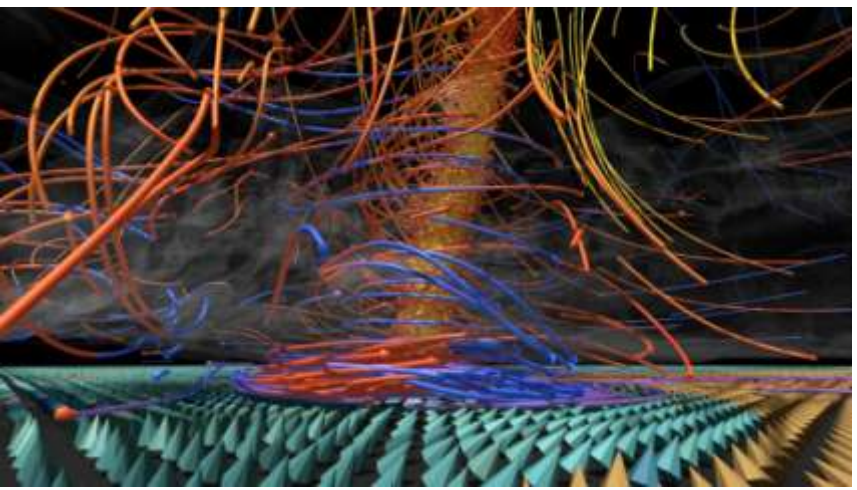
Cisco Has Built 10 GigE Waves on CENIC, PW, & NLR and Installed Large 6506 Switches for Access Points in San Diego, Los Angeles, Sunnyvale, Seattle, Chicago and McLean for CineGrid Members

Some of These Points are also GLIF GOLEs

*CTWatch QUARTERLY May 2007

CineGrid 4K Digital Cinema Projects: "Learning by Doing"

CineGrid @ iGrid 2005



CineGrid @ AES 2006





CineGrid @ Holland Festival 2007



CineGrid @ GLIF 2007

Laurin Herr, Pacific Interface; Tom DeFanti, Calit2

First Tri-Continental Premier of a Streamed 4K Feature Film With Global HD Discussion



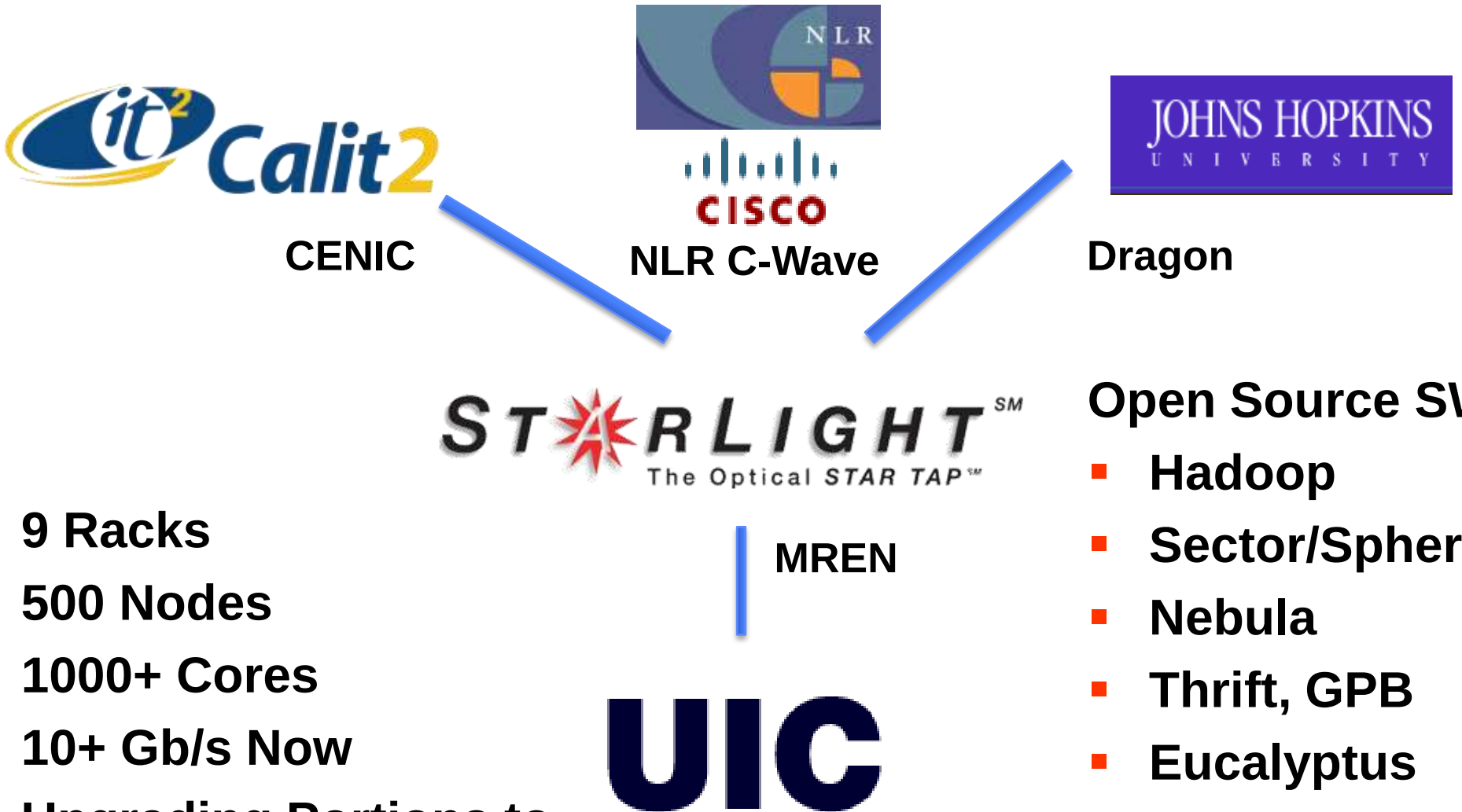
4K Transmission Over 10Gbps--
4 HD Projections from One 4K Projector

CineGrid 4K Remote Microscopy Collaboratory: USC to Calit2



Richard Weinberg, USC

Open Cloud OptIPuter Testbed--Manage and Compute Large Datasets Over 10Gbps Lambdas



- 9 Racks
- 500 Nodes
- 1000+ Cores
- 10+ Gb/s Now
- Upgrading Portions to 100 Gb/s in 2010/2011

Open Source SW

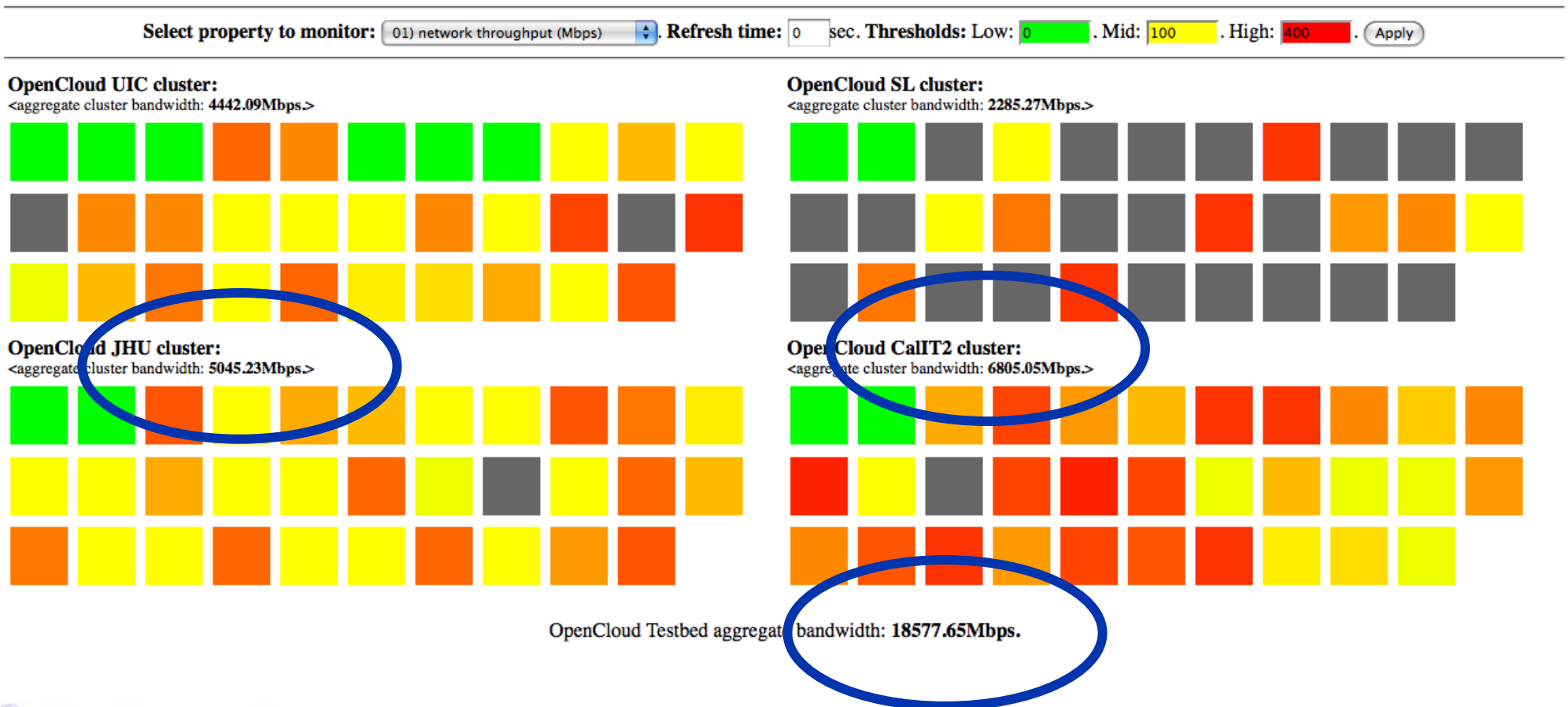
- Hadoop
- Sector/Sphere
- Nebula
- Thrift, GPB
- Eucalyptus
- Benchmarks

Open Cloud Consortium

Source: Robert Grossman, UChicago

Terasort on Open Cloud Testbed Sustains >5 Gbps--Only 5% Distance Penalty!

Sorting 10 Billion Records (1.2 TB) at 4 Sites (120 Nodes)



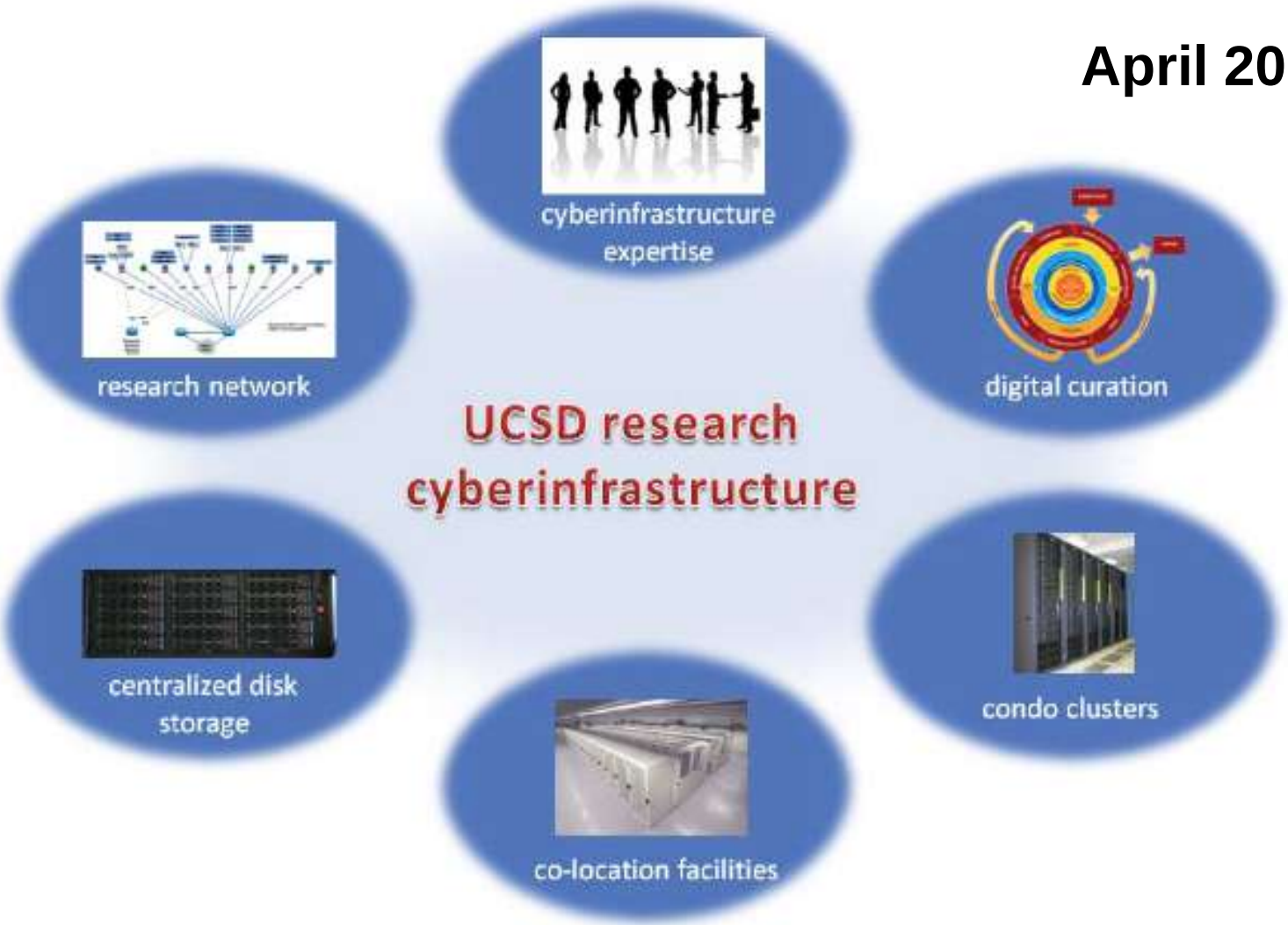
Source: Robert Grossman, UChicago

"Blueprint for the Digital University"--Report of the UCSD Research Cyberinfrastructure Design Team

- **Focus on Data-Intensive Cyberinfrastructure**

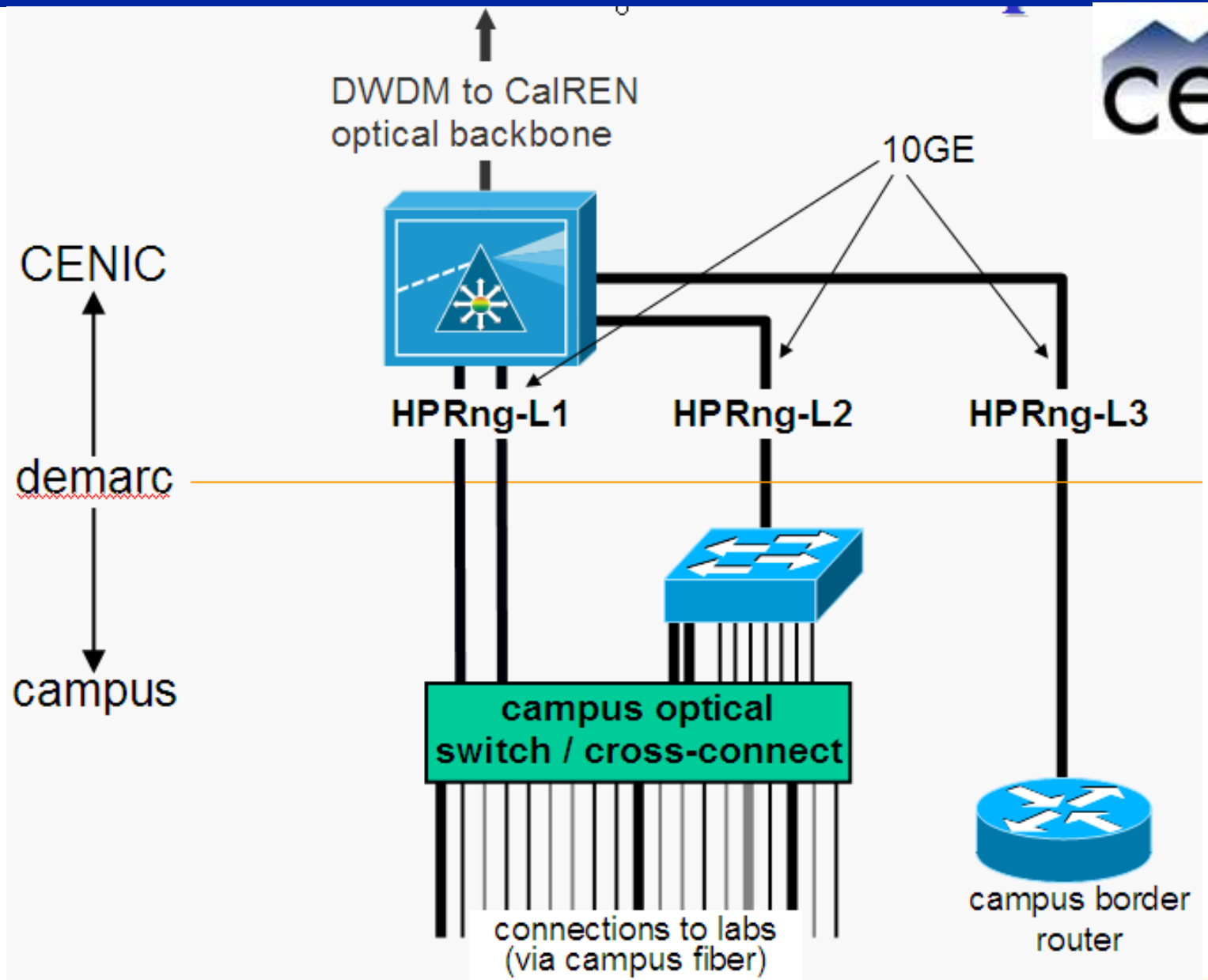
April 2009

No Data Bottlenecks --Design for Gigabit/s Data Flows



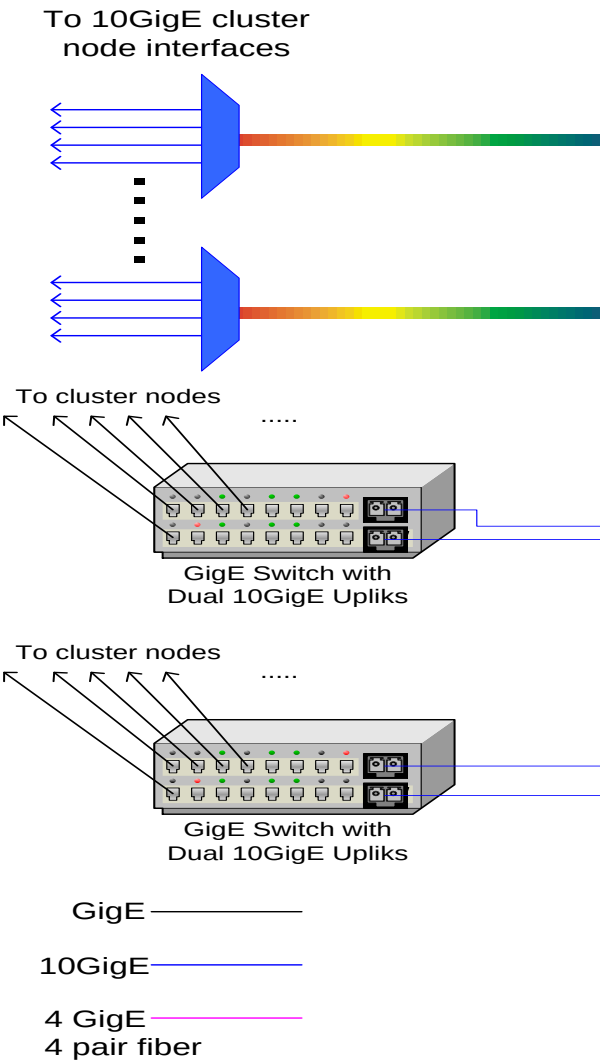
research.ucsd.edu/documents/rcidt/RCIDTReportFinal2009.pdf

Campus Preparations Needed to Accept CENIC CalREN Handoff to Campus



Source: Jim Dolgonas, CENIC

Current UCSD Prototype Optical Core: Bridging End-Users to CENIC L1, L2, L3 Services



Enpoints:

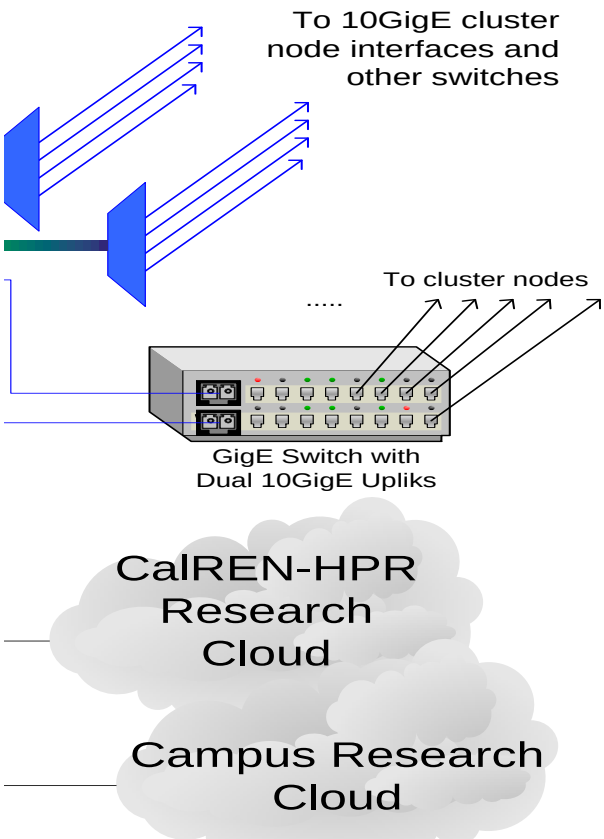
>= 60 endpoints at 10 GigE

>= 32 Packet switched

>= 32 Switched wavelengths

>= 300 Connected endpoints

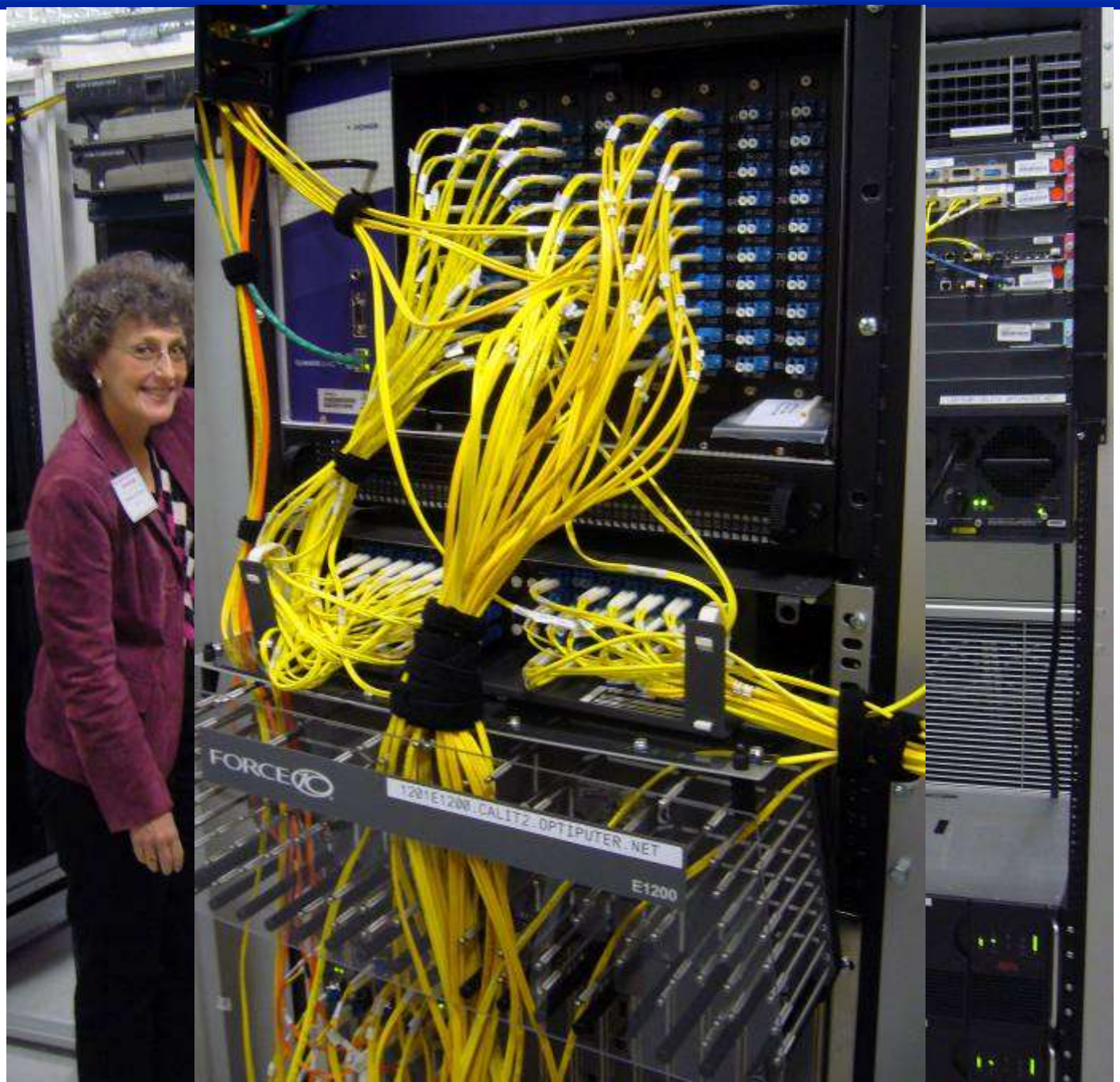
Approximately 0.5 TBit/s Arrive at the "Optical" Center of Campus. Switching is a Hybrid of: Packet, Lambda, Circuit -- OOO and Packet Switches



Source: Phil Papadopoulos, SDSC/Calit2 (Quartzite PI, OptIPuter co-PI)

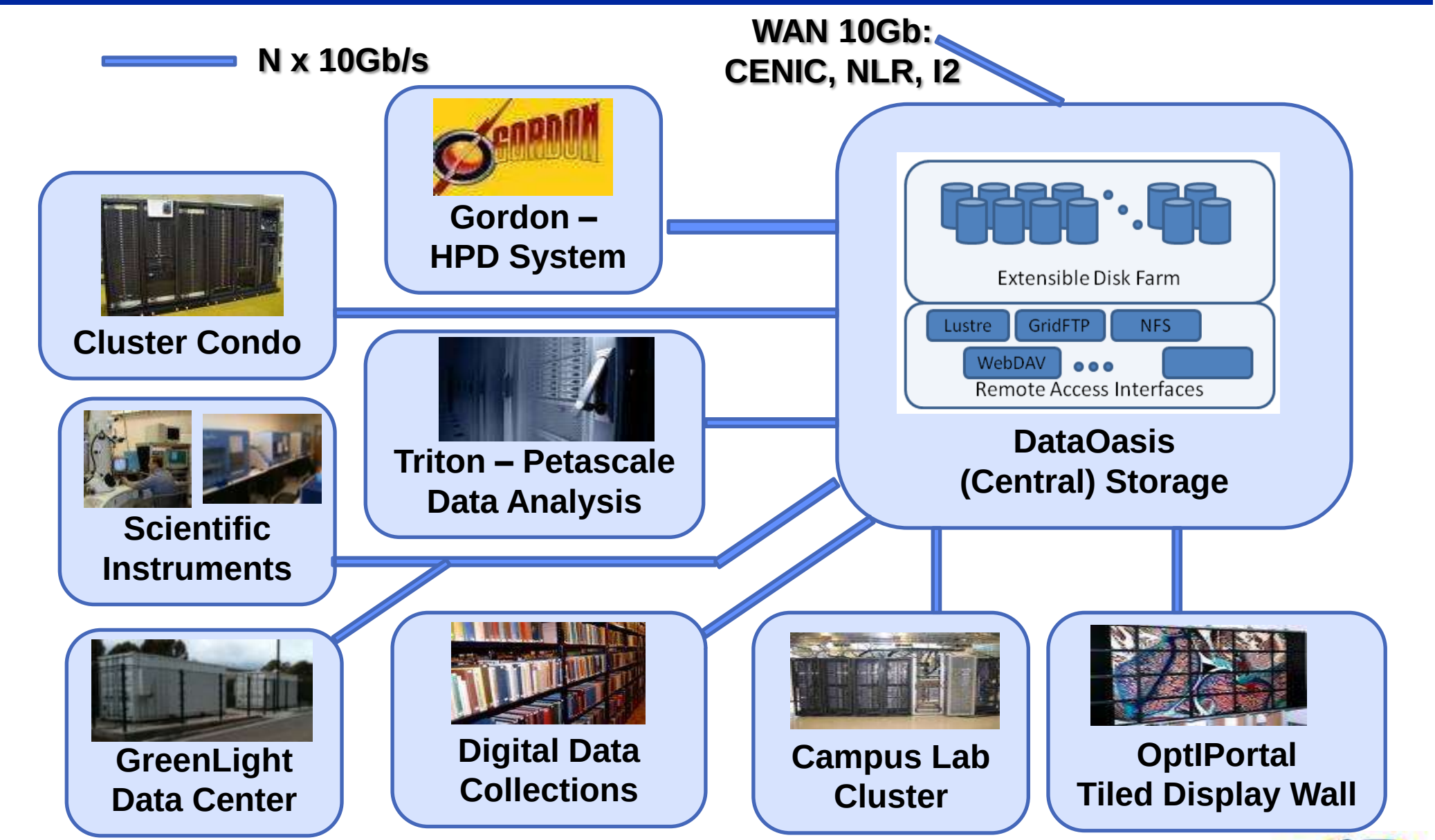
Quartzite Network MRI #CNS-0421555; OptIPuter #ANI-0225642

Calit2 Sunlight Optical Exchange Contains Quartzite



Maxine Brown, EVL, UIC OptIPuter Project Manager

UCSD Campus Investment in Fiber Enables Consolidation of Energy Efficient Computing & Storage



Source: Philip Papadopoulos, SDSC, UCSD

National Center for Microscopy and Imaging Research: Integrated Infrastructure of Shared Resources

NCMIR



Shared Infrastructure

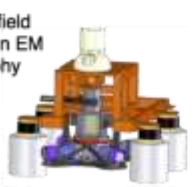


Scientific Instruments

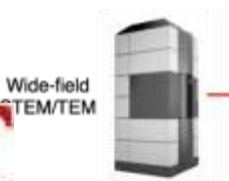




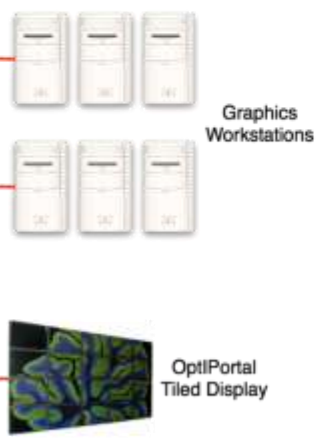




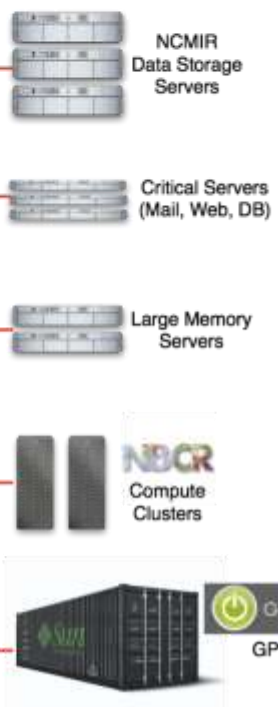




End User Workstations



Local SOM Infrastructure



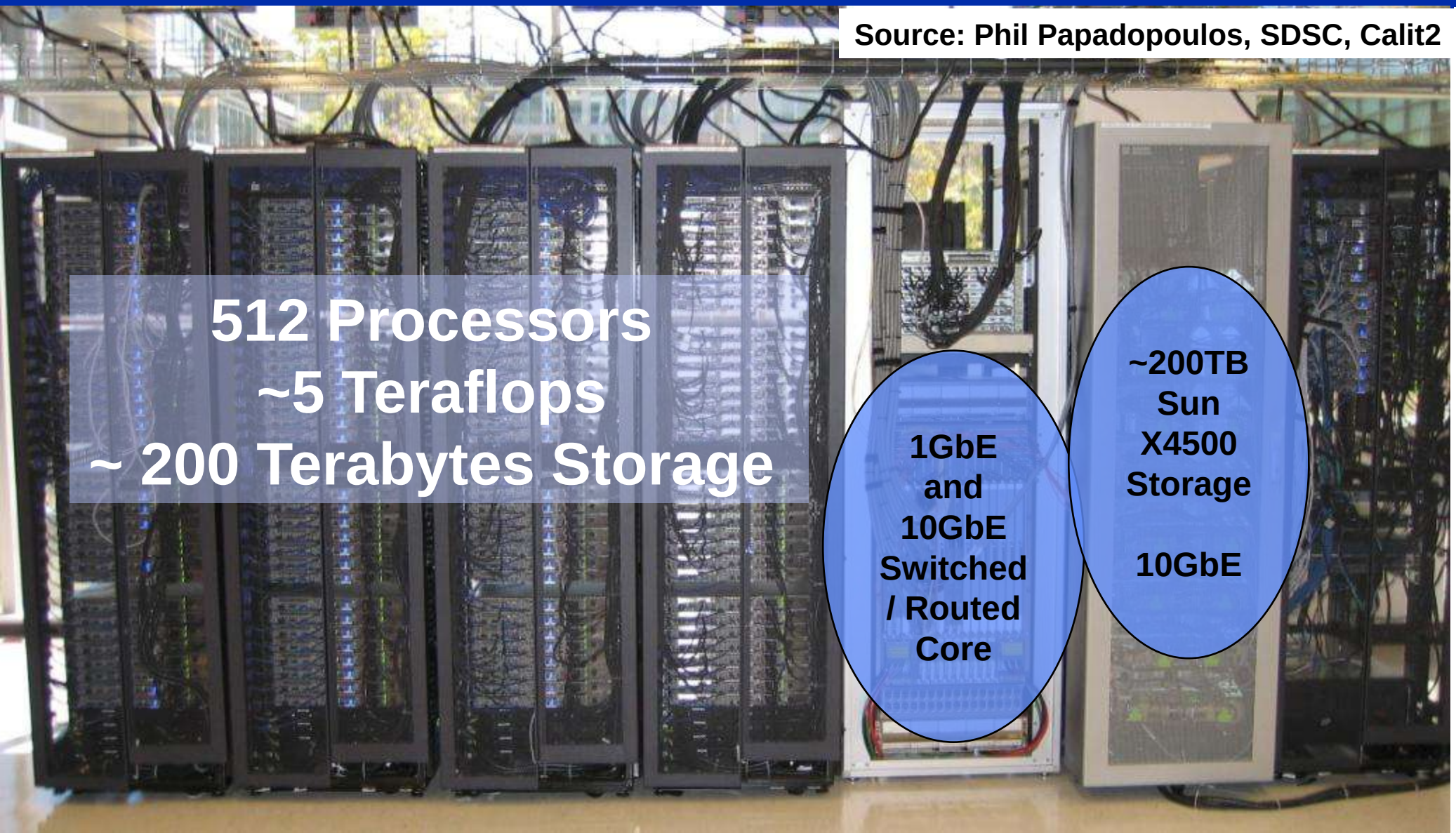
SDSC

Source: Steve Peltier, NCMIR

Community Cyberinfrastructure for Advanced Microbial Ecology Research and Analysis



Calit2 Microbial Metagenomics Cluster-Lambda Direct Connect Science Data Server



4000 Users
From 90 Countries

Gordon and Betty MOORE FOUNDATION

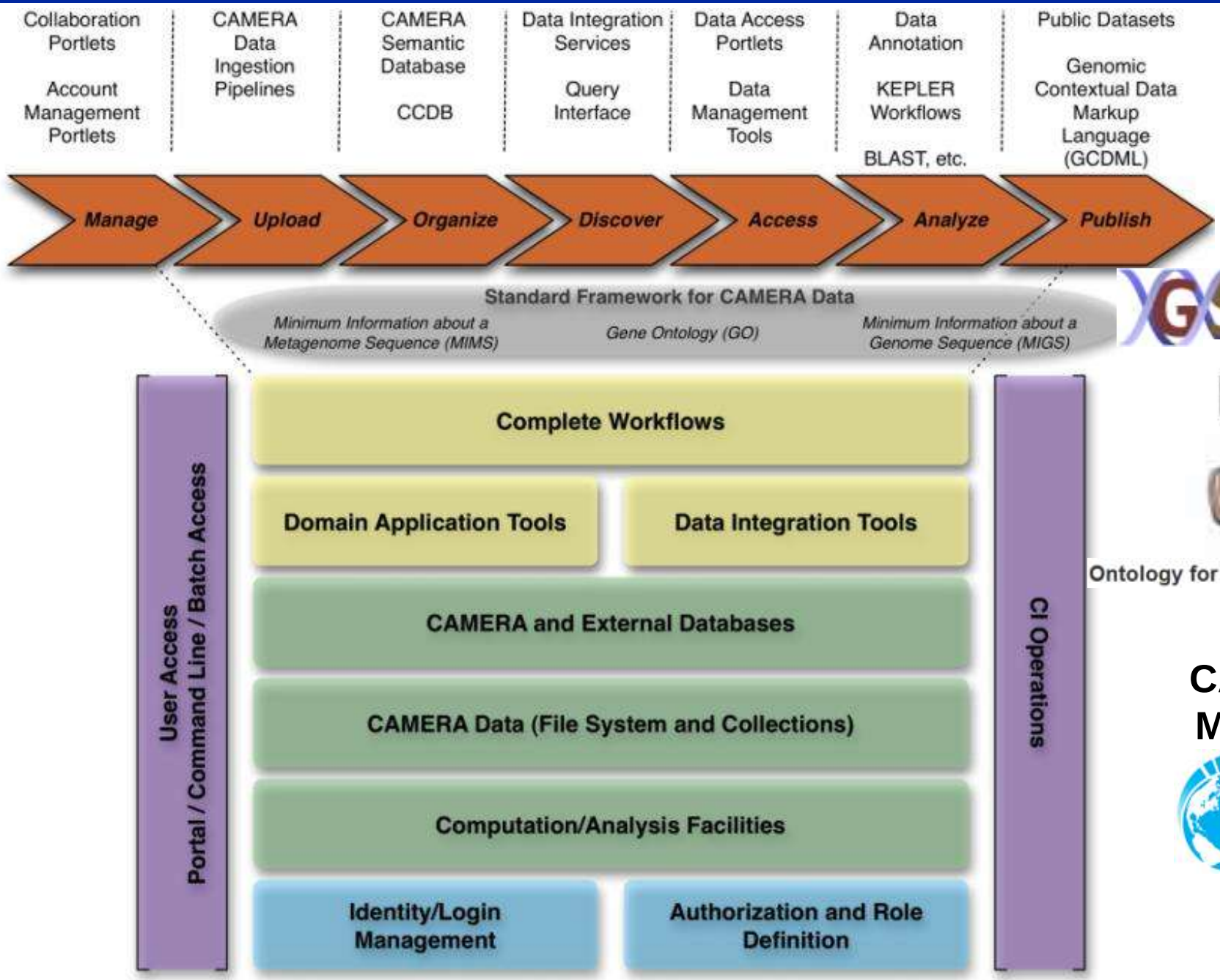
ROCKS

SDSC

J. Craig Venter INSTITUTE

it²

Creating CAMERA 2.0 -
Advanced Cyberinfrastructure Service Oriented Architecture



GSC Genomic Standards Consortium

Env

OBI

Ontology for Biomedical Investigations

Source:
CAMERA CTO
Mark Ellisman

camera
Marine Microbial Ecology

it²

OptIPuter Persistent Infrastructure Enables Calit2 and U Washington CAMERA Collaboratory



OptIPuter

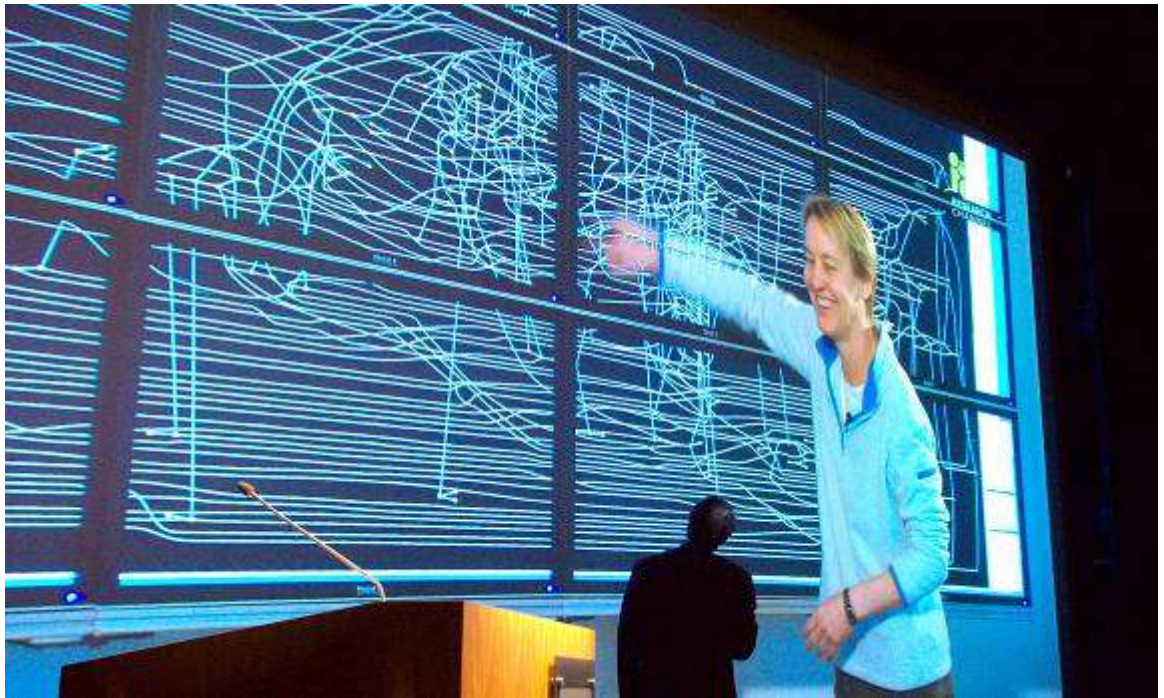
Ginger Armbrust's Diatoms: Micrographs, Chromosomes, Genetic Assembly

ROCKS



Gordon and Betty MOORE FOUNDATION

camera
Marine Microbial Ecology



NSF

it²

NSF Funds a Data-Intensive Track 2 Supercomputer: SDSC's Gordon-Coming Summer 2011

- **Data-Intensive Supercomputer Based on SSD Flash Memory and Virtual Shared Memory SW**
  - *Emphasizes MEM and IOPS over FLOPS*
  - *Supernode has Virtual Shared Memory:*
    - *2 TB RAM Aggregate*
    - *8 TB SSD Aggregate*
    - *Total Machine = 32 Supernodes*
    - *4 PB Disk Parallel File System >100 GB/s I/O*



- **System Designed to Accelerate Access to Massive Data Bases being Generated in Many Fields of Science, Engineering, Medicine, and Social Science**

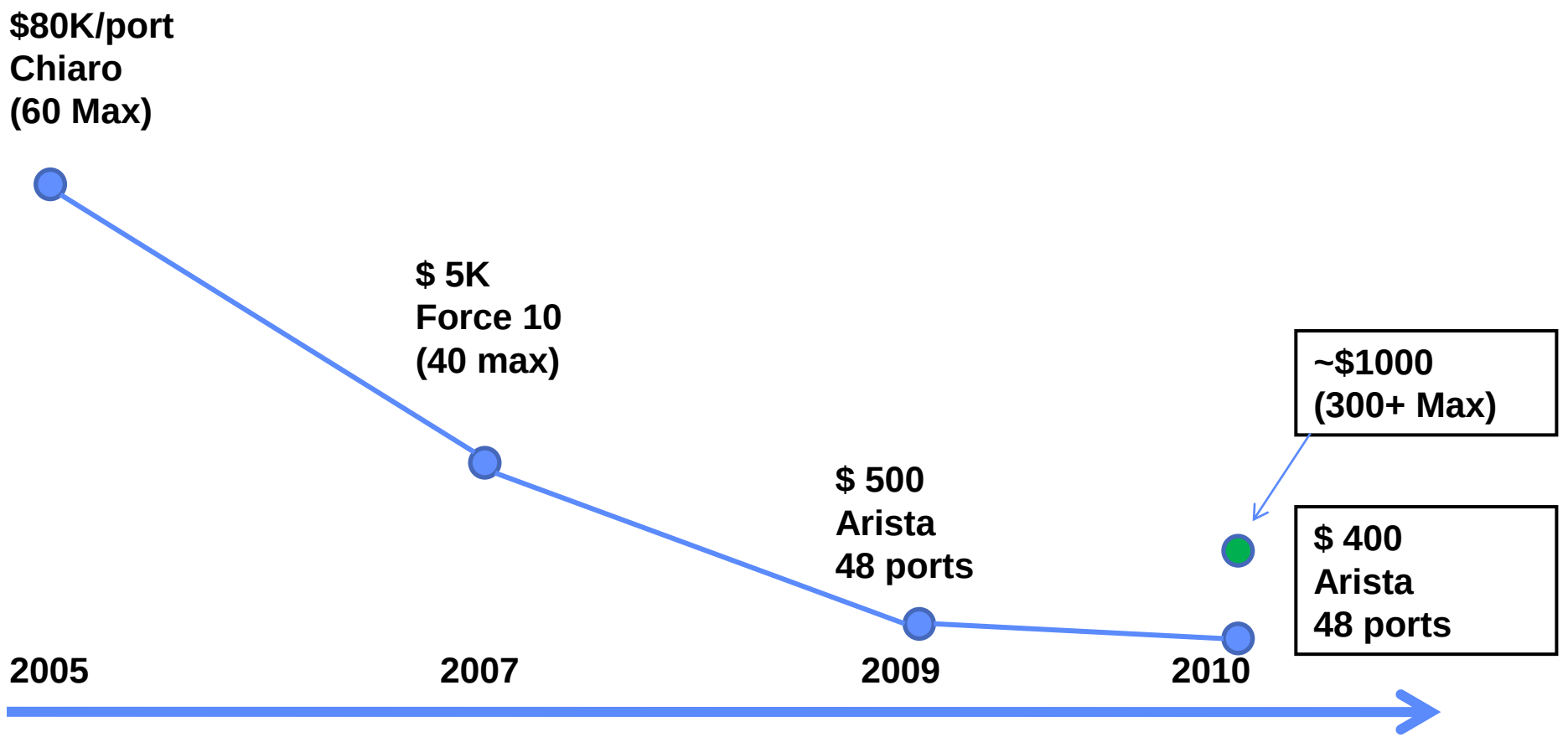
SDSC

Source: Mike Norman, Allan Snavely SDSC

it²

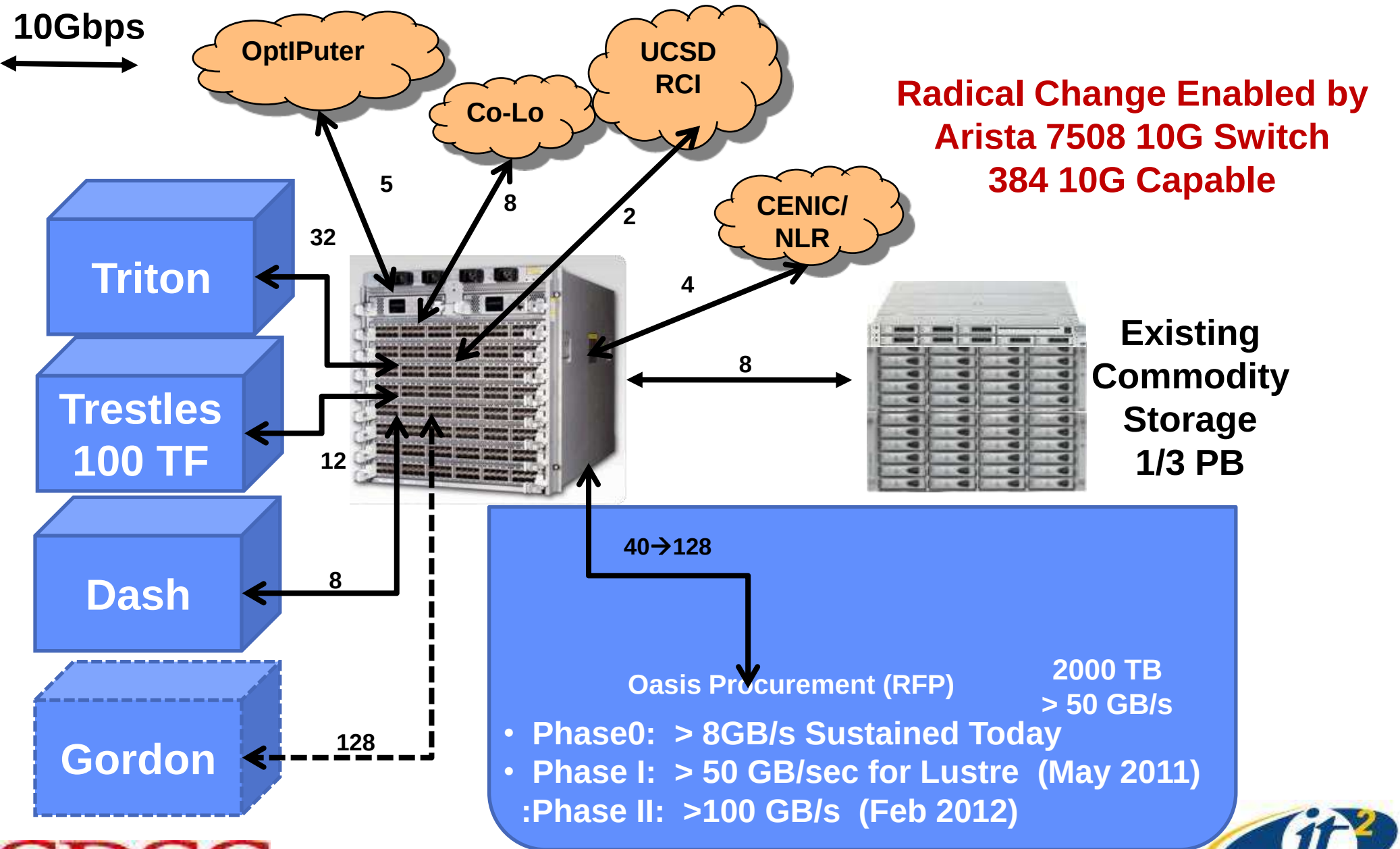
Rapid Evolution of 10GbE Port Prices Makes Campus-Scale 10Gbps CI Affordable

- **Port Pricing is Falling**
- **Density is Rising – Dramatically**
- **Cost of 10GbE Approaching Cluster HPC Interconnects**



Source: Philip Papadopoulos, SDSC/Calit2

Arista Enables SDSC's Massive Parallel 10G Switched Data Analysis Resource



Radical Change Enabled by Arista 7508 10G Switch 384 10G Capable

SDSC

Source: Philip Papadopoulos, SDSC/Calit2

it²

Data Oasis – 3 Different Types of Storage



HPC Storage (Lustre-Based PFS)

- Purpose: Transient Storage to Support HPC, HPD, and Visualization
- Access Mechanisms: Lustre Parallel File System Client



Project (Traditional File Server) Storage

- Purpose: Typical Project / User Storage Needs
- Access Mechanisms: NFS/CIFS "Network Drives"



Cloud Storage

- Purpose: Long-Term Storage of Data that will be Infrequently Accessed
- Access Mechanisms: S3 interfaces, DropBox-esq web interface, CommVault

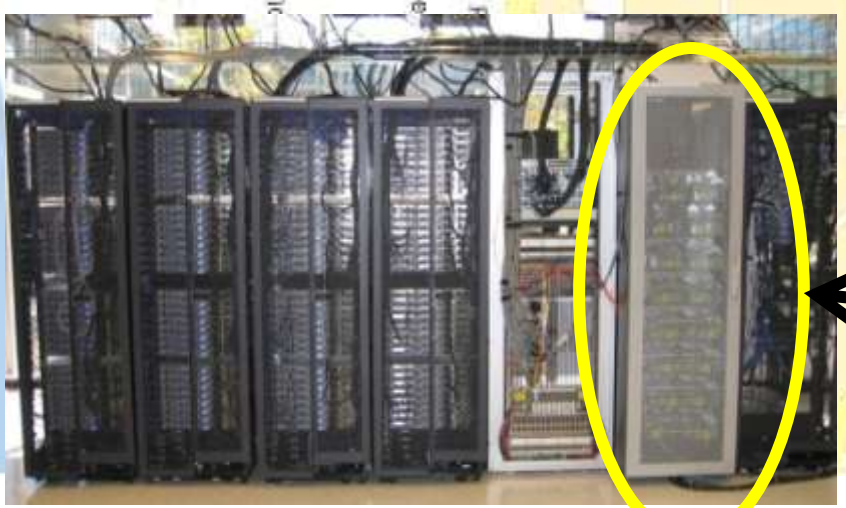
SDSC

it²

Calit2 CAMERA Automatic Overflows into SDSC Triton



@ CALIT2



@ SDSC

Triton Resource

CAMERA - Managed Job Submit Portal (VM)

Transparently Sends Jobs to Submit Portal on Triton

10Gbps

CAMERA DATA

Direct Mount == No Data Staging

SDSC

ROCKS

California and Washington Universities Are Testing a 10Gbps Lambda-Connected Commercial Data Cloud

- **Amazon Experiment for Big Data**
  - **Only Available Through CENIC & Pacific NW GigaPOP**
    - **Private 10Gbps Peering Paths**
  - **Includes Amazon EC2 Computing & S3 Storage Services**
- **Early Experiments Underway**
  - **Phil Papadopoulos, Calit2/SDSC Rocks**
  - **Robert Grossman, Open Cloud Consortium**

SDSC

amazon web services

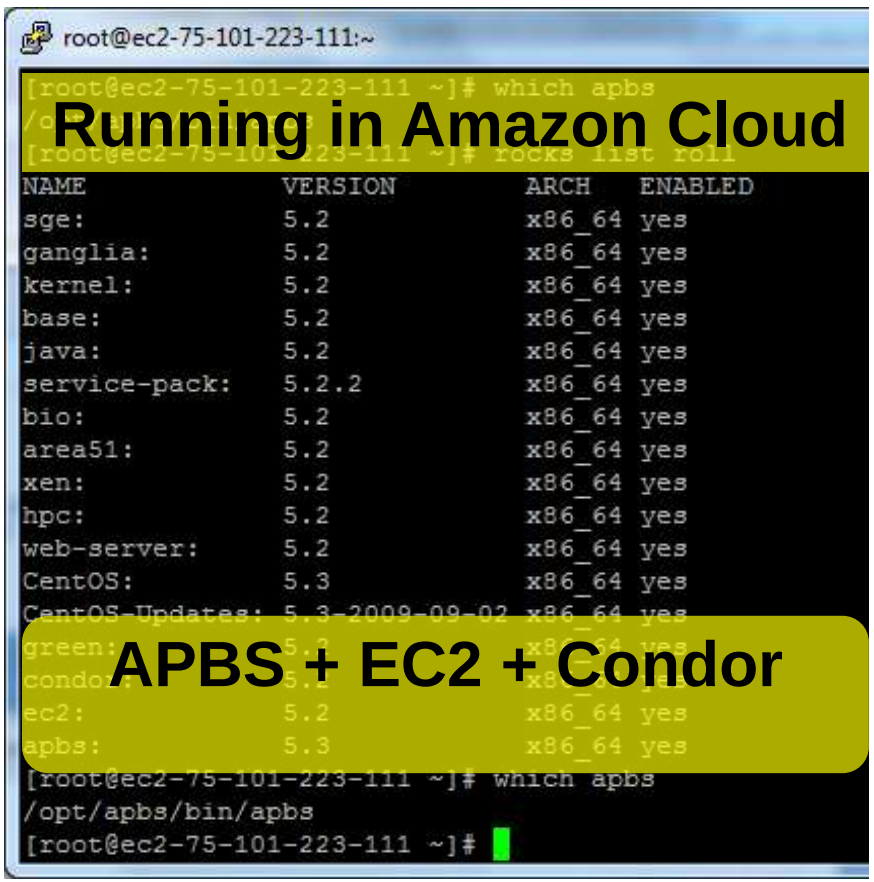
cenic

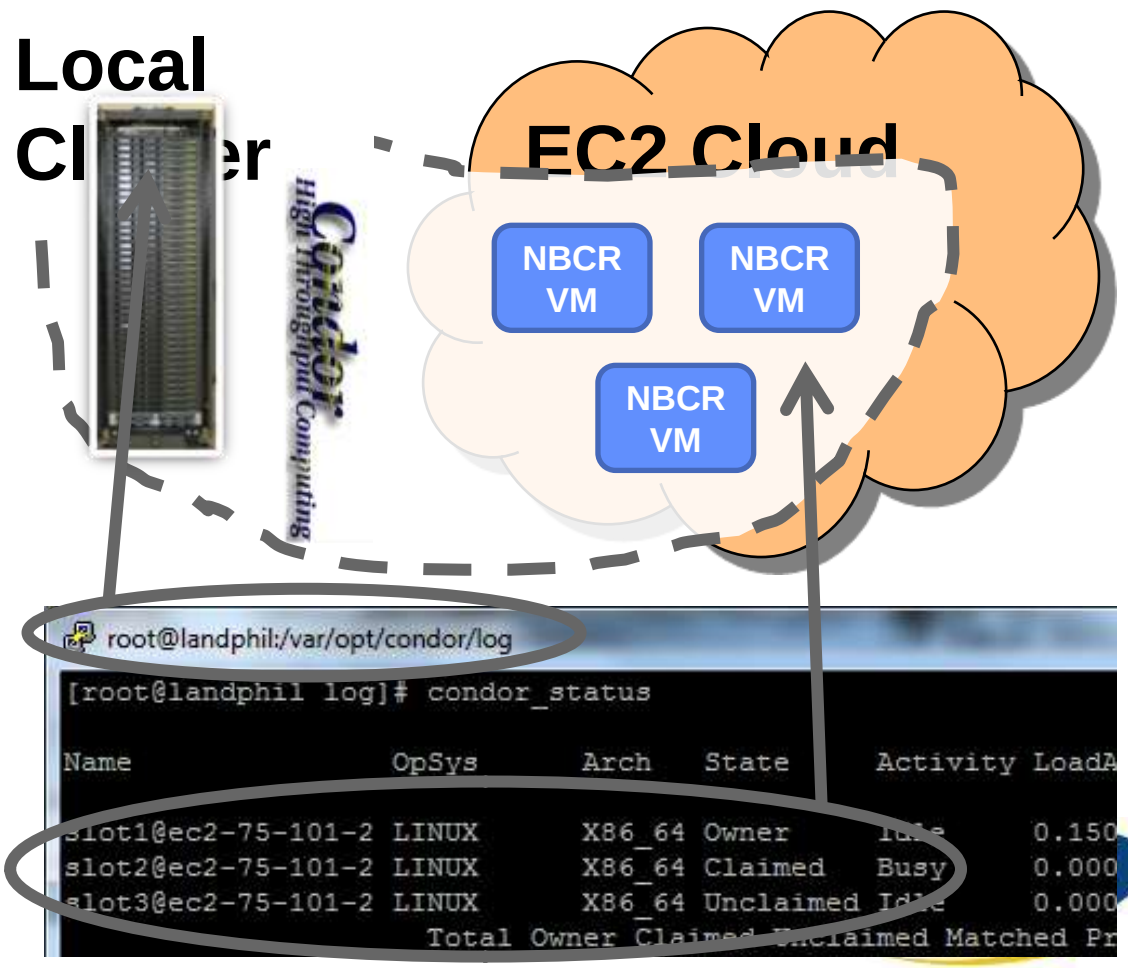
PACIFIC NORTHWEST GIGAPOP

it²

Using Condor and Amazon EC2 on Adaptive Poisson-Boltzmann Solver (APBS)

- APBS Rocks Roll (NBCR) + EC2 Roll + Condor Roll = Amazon VM
- Cluster extension into Amazon using Condor





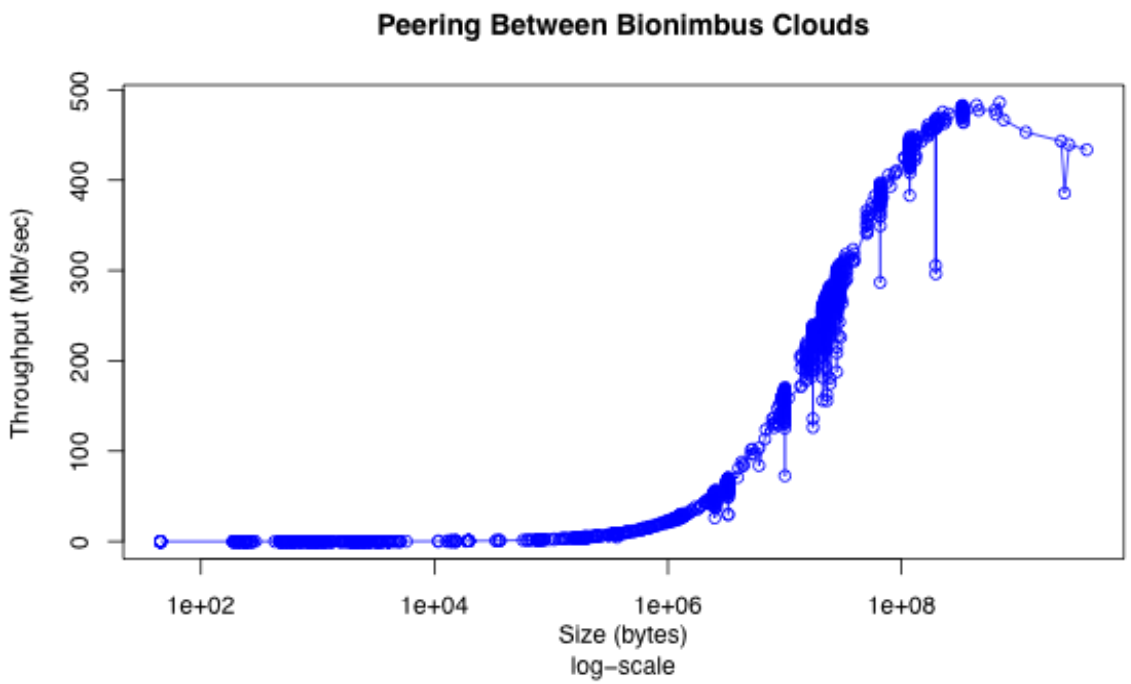
SDSC

Source: Phil Papadopoulos, SDSC/Calit2

Hybrid Cloud Computing with modENCODE Data

- Computations in Bionimbus Can Span the Community Cloud & the Amazon Public Cloud to Form a Hybrid Cloud
- Sector was used to Support the Data Transfer between Two Virtual Machines
  - One VM was at UIC and One VM was an Amazon EC2 Instance
- Graph Illustrates How the Throughput between Two Virtual Machines in a Wide Area Cloud Depends upon the File Size



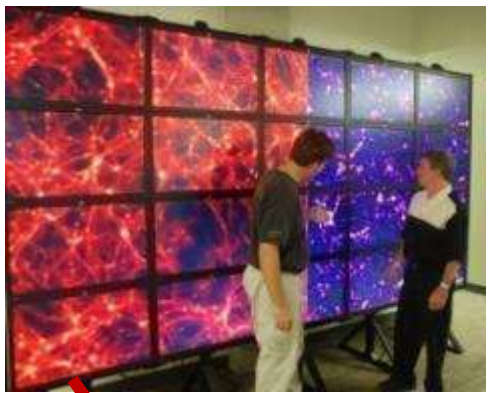




Biological data (Bionimbus)

Source: Robert Grossman, UChicago

OptIPlanet Collaboratory: Enabled by 10Gbps "End-to-End" Lightpaths

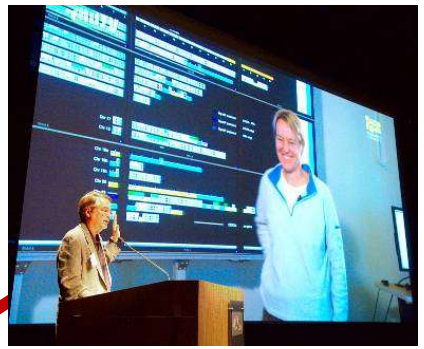


End User OptIPortal



Local or Remote Instruments





HD/4k Live Video

HPC

TeraGrid





10G Lightpaths

Campus Optical Switch

amazon web services



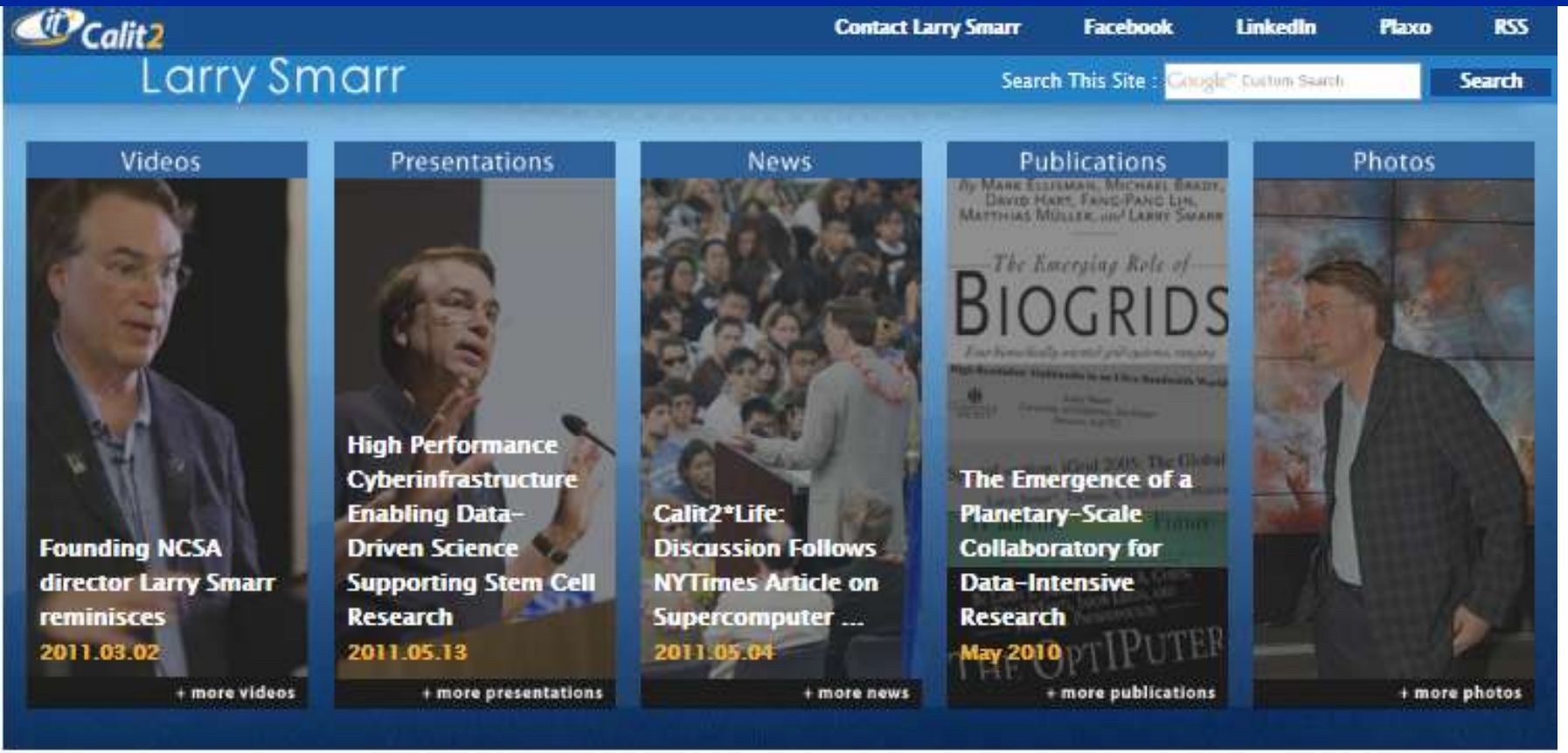
Data Repositories & Clusters



HD/4k Video Repositories

it²

You Can Download This Presentation at lsmarr.calit2.net



ABOUT LARRY SMARR

Larry Smarr is the founding Director of the California Institute for Telecommunications and Information Technology (Calit2), a UC San Diego/UC Irvine partnership, and holds the Harry E. Gruber professorship in Computer Science and Engineering (CSE) at UCSD's Jacobs School. At Calit2, Smarr has continued to drive major developments in information infrastructure-- including the Internet,

WHAT I'M READING

Complete List of Larry's Recommendations

Homo Britannicus: The Incredible Story of Human Life in Britain (2006)

American College and University Presidents' Climate Commitment

A Doctor's Vision of the Future of Medicine

RECENT TWITTER POSTS

Twitter? Click To Follow Me

lsmarr: Light going up-what an energy waste! RT @ISS_Research: Night view of US East Coast from ISS Crew Earth Observations http://1.usa.gov/INWw5d *Tue, 24 May*